

THE UNIVERSITY OF CHICAGO

ANTECEDENCE AND CONTEXT: IDENTITY AND NONIDENTITY IN VERB
PHRASE ELLIPSIS AND ANAPHORIC DEACCENTING

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE HUMANITIES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF LINGUISTICS

BY
JEFFREY PATRICK GEIGER

CHICAGO, ILLINOIS

AUGUST 2020

Copyright © 2020 by Jeffrey Patrick Geiger
All Rights Reserved

For Mom

TABLE OF CONTENTS

LIST OF FIGURES	x
LIST OF TABLES	xiii
ACKNOWLEDGMENTS	xiv
ABSTRACT	xvii
1 INTRODUCTION	1
2 BACKGROUND: IDENTITY AND NONIDENTITY IN VERB PHRASE ELLIPSIS	13
2.1 Evidence for syntactic identity	15
2.2 Evidence for semantic identity	21
2.3 Resolving overgeneration	25
2.4 Resolving undergeneration	30
2.5 Antecedentless VP ellipsis	38
2.6 An empirical gap	44
3 VERB PHRASE ELLIPSIS INTERPRETATION IN COMPLEX DISCOURSE CON- TEXTS	50
3.1 Experiment 1: VPE interpretation in context	53
3.1.1 Design and materials	53
3.1.2 Procedure	58
3.1.3 Participants	59
3.1.4 Predictions	60
3.1.5 Results	62
3.1.6 Analysis	63
3.1.7 Discussion	65
3.2 Experiment 2: Ellipsis-specific context effects	67
3.2.1 Design, materials, and procedure	69
3.2.2 Participants	69
3.2.3 Predictions	71
3.2.4 Results	71
3.2.5 Analysis	73
3.2.6 Discussion	74
3.3 Experiment 3: Discourse availability of VPE interpretations	76
3.3.1 Design, materials, and procedure	77
3.3.2 Participants	77
3.3.3 Predictions	79
3.3.4 Results	80
3.3.5 Analysis	81
3.3.6 Discussion	83
3.4 Experiment 4: Interpretation of <i>do that</i> in context	84

3.4.1	Design, materials, and procedure	86
3.4.2	Participants	87
3.4.3	Predictions	87
3.4.4	Results	89
3.4.5	Analysis	90
3.4.6	Discussion	92
3.5	General discussion	94
3.5.1	VPE interpretation in context	95
3.5.2	Comparing VPE and <i>do that</i>	103
4	A PROBABILISTIC MODEL OF VPE INTERPRETATION IN CONTEXT . . .	109
4.1	Experiments 5 and 6: Empirical priors – Availability of competing meanings before VPE	110
4.1.1	Design, methods, and materials	111
4.1.2	Participants	113
4.1.3	Results	115
4.1.4	Analysis	117
4.1.5	Discussion	119
4.2	Experiment 7: Empirical posteriors – Forced-choice VPE interpretation . . .	119
4.2.1	Design, methods, and procedure	120
4.2.2	Participants	120
4.2.3	Results	121
4.2.4	Analysis	123
4.2.5	Discussion	124
4.3	Probabilistic models of VPE interpretation in context	126
4.3.1	Conceptual and terminological overview	127
4.3.2	Identity-only model	130
4.3.3	Identity model with error	138
4.3.4	Salience-only model	143
4.3.5	Salience model with error	146
4.3.6	Hybrid model (no error)	148
4.3.7	Hybrid model with error	151
4.3.8	Model comparison	153
4.3.9	Discussion	164
4.4	General discussion	166
5	BACKGROUND: GIVENNESS, INFERABILITY, AND DEACCENTING	172
5.1	Diagnosing emphasis	180
5.1.1	Correlates of accent and correlates of stress	181
5.1.2	Complications of prosodic realization	186
5.2	Theories of deaccenting licensing	189
5.2.1	The grammatical approach	190
5.2.2	The accommodation approach	223
5.2.3	Summary	235

6	PRODUCTION OF DEACCENTING UNDER IDENTITY AND INFERENCE .	239
6.1	Experiment 8: Norming verb inferability	244
6.1.1	Design and materials	245
6.1.2	Procedure	247
6.1.3	Participants	249
6.1.4	Results and stimulus selection	249
6.2	Experiment 9: Production of deaccenting	250
6.2.1	Design and materials	251
6.2.2	Procedure	255
6.2.3	Participants	257
6.2.4	Data processing	257
6.2.5	Predictions	261
6.2.6	Results	263
6.2.7	Analysis	264
6.2.8	Discussion	276
6.3	Experiment 10: Perception of emphasis in production data	284
6.3.1	Design and materials	286
6.3.2	Procedure	287
6.3.3	Participants	289
6.3.4	Predictions	290
6.3.5	Results	290
6.3.6	Analysis	291
6.3.7	Discussion	295
6.4	General discussion	297
7	PERCEPTION OF DEACCENTING UNDER IDENTITY AND INFERENCE .	301
7.1	Experiment 11: Perception of emphasis in out-of-the-blue contexts	304
7.1.1	Design and materials	305
7.1.2	Procedure	307
7.1.3	Participants	310
7.1.4	Predictions	310
7.1.5	Results	313
7.1.6	Analysis	316
7.1.7	Discussion	317
7.2	Experiments 12 and 13: Perception of emphasis in supportive contexts . . .	322
7.2.1	Design, materials, and procedure	324
7.2.2	Participants	327
7.2.3	Predictions	327
7.2.4	Results	328
7.2.5	Analysis	334
7.2.6	Discussion	336
7.3	Experiment 14: Perception of emphasis with presupposition triggers	341
7.3.1	Design, materials, and procedure	342
7.3.2	Participants	343
7.3.3	Predictions	343

7.3.4	Results	344
7.3.5	Analysis	347
7.3.6	Discussion	349
7.4	General discussion	351
8	CONCLUSION	356
8.1	Summary of findings: Verb phrase ellipsis	356
8.2	Summary of findings: Deaccenting	364
8.3	Synthesis, discussion, and future directions	372
A	EXPERIMENT 1: VPE INTERPRETATION IN CONTEXT	379
A.1	Task and layout	379
A.2	Participant statistics and data loss	380
A.3	Critical items	380
A.4	By-item results	386
A.5	Fillers	386
B	EXPERIMENT 2: ELLIPSIS-SPECIFIC INTERPRETATION EFFECTS	392
B.1	Task and layout	392
B.2	Participant statistics and data loss	393
B.3	Critical items	393
B.4	By-item results	396
B.5	Fillers	396
C	EXPERIMENT 3: DISCOURSE COHERENCE OF VPE INTERPRETATIONS	400
C.1	Task and layout	400
C.2	Participant statistics and data loss	400
C.3	Critical items	401
C.4	By-item results	403
C.5	Fillers	403
D	EXPERIMENT 4: INTERPRETATION OF <i>DO THAT</i> IN CONTEXT	407
D.1	Task and layout	407
D.2	Participant statistics and data loss	408
D.3	Critical items	408
D.4	By-item results	411
D.5	Fillers	411
E	EXPERIMENT 5: EMPIRICAL PRIORS, THREE MESSAGES	415
E.1	Task and layout	415
E.2	Participant statistics and data loss	416
E.3	Critical items	416
E.4	Fillers – Not analyzed	418
E.5	Fillers – Attention screening	419

F	EXPERIMENT 6: EMPIRICAL PRIORS, TWO MESSAGES	421
F.1	Task and layout	421
F.2	Participant statistics and data loss	422
F.3	Critical items	422
F.4	Fillers	422
G	EXPERIMENT 7: FORCED-CHOICE VPE INTERPRETATION (EMPIRICAL POSTERiors)	423
G.1	Task and layout	423
G.2	Participant statistics and data loss	424
G.3	Critical items	424
G.4	Fillers – Not analyzed	426
G.5	Fillers – Attention screening	429
H	MODELING CALCULATIONS AND MODEL COMPARISON	431
H.1	Sample model calculations	431
H.2	Bayes factors comparing identity and salience models	435
I	EXPERIMENT 8: NORMING VERB INFERABILITY	437
I.1	Note on “new” verbs	437
I.2	Task and layout	438
I.3	Participant statistics and data loss	438
I.4	Aggregate results	439
I.5	By-item results	440
J	EXPERIMENT 9: PRODUCTION OF DEACCENTING	442
J.1	Participant statistics and data loss	442
J.2	Materials	443
J.3	Results and analysis for relativized duration	446
K	EXPERIMENT 10: PERCEPTION OF EMPHASIS IN PRODUCTION DATA	451
K.1	Task and layout	451
K.2	Participant statistics and data loss	452
K.3	Materials	452
L	EXPERIMENT 11: OUT-OF-THE-BLUE DEACCENTING PERCEPTION	453
L.1	Task and layout	453
L.2	Participant statistics and data loss	453
L.3	Additional recording stimuli	454
L.4	Critical stimuli	464
L.5	Fillers	465
M	EXPERIMENT 12: DEACCENTING PERCEPTION IN CONTEXT, I	466
M.1	Task and layout	466
M.2	Participant statistics and data loss	466
M.3	Critical items	466

M.4	Fillers	468
N	EXPERIMENT 13: DEACCENTING PERCEPTION IN CONTEXT, II	470
N.1	Task and layout	470
N.2	Participant statistics and data loss	470
N.3	Materials	471
O	EXPERIMENT 14: DEACCENTING PERCEPTION WITH PRESUPPOSITION TRIGGERS	472
O.1	Task and layout	472
O.2	Participant statistics and data loss	472
O.3	Materials	473
	REFERENCES	474

LIST OF FIGURES

3.1	Sample screen from Experiment 1	54
3.2	Experiment 1 results	62
3.3	Experiment 2 results	72
3.4	Experiment 3 results	80
3.5	Experiment 4 results	90
4.1	Sample screen from Experiment 5	114
4.2	Experiment 5 results	116
4.3	Experiment 5 results with zero interpretation removed	116
4.4	Experiment 6 results	117
4.5	Sample screen from Experiment 7.	121
4.6	Experiment 7 results	122
4.7	Predicted vs. empirical posteriors: identity model (no error)	137
4.8	Predicted vs. empirical posteriors: identity model with error	141
4.9	Predicted vs. empirical posteriors: salience model (no error)	145
4.10	Predicted vs. empirical posteriors: salience model with error	147
4.11	Predicted vs. empirical posteriors: hybrid model (no error)	150
4.12	Predicted vs. empirical posteriors: hybrid model with error	152
6.1	Sample screen from Experiment 8	248
6.2	Experiment 9 results: z-transformed intensity	265
6.3	Experiment 9 results: z-transformed f_0	265
6.4	Experiment 9 results: Duration	266
6.5	Experiment 9 results: z-transformed intensity, O1	267
6.6	Experiment 9 results: z-transformed f_0 , O1	268
6.7	Experiment 9 results: Duration, O1	269
6.8	Experiment 9 results: z-transformed intensity, O2	270
6.9	Experiment 9 results: z-transformed f_0 , O2	270
6.10	Experiment 9 results: Duration, O2	271
6.11	Experiment 9 results: z-transformed intensity, V1	272
6.12	Experiment 9 results: z-transformed f_0 , V1	272
6.13	Experiment 9 results: Duration, V1	274
6.14	Experiment 9 results: z-transformed intensity, V2	274
6.15	Experiment 9 results: z-transformed f_0 , V2	275
6.16	Experiment 9 results: Duration, V2	276
6.17	Experiment 9 results: z-transformed intensity, S	277
6.18	Experiment 9 results: z-transformed f_0 , S	277
6.19	Experiment 9 results: Duration, S	278
6.20	Sample screens from Experiment 10	288
6.21	Experiment 10 results: Objects	292
6.22	Experiment 10 results: Verbs	292
7.1	Sample screens from Experiment 11	309
7.2	Experiment 11 results: Entailment items by verb status	314

7.3	Experiment 11 results: Bridging items by verb status	314
7.4	Experiment 11 results: Entailment items by accent status	315
7.5	Experiment 11 results: Bridging items by accent status	315
7.6	Sample screens from Experiment 12	324
7.7	Sample screens from Experiment 13	325
7.8	Experiment 12 results: Entailment items by verb status	329
7.9	Experiment 12 results: Bridging items by verb status	329
7.10	Experiment 12 results: Entailment items by accent status	330
7.11	Experiment 12 results: Bridging items by accent status	330
7.12	Experiment 13 results: Entailment items by verb status	332
7.13	Experiment 13 results: Bridging items by verb status	332
7.14	Experiment 13 results: Entailment items by accent status	333
7.15	Experiment 13 results: Bridging items by accent status	333
7.16	Experiment 14 results: Entailment items by verb status	345
7.17	Experiment 14 results: Bridging items by verb status	345
7.18	Experiment 14 results: Entailment items by accent status	346
7.19	Experiment 14 results: Bridging items by accent status	346
A.1	Sample screen from Experiment 1	379
A.2	Experiment 1 results by item	386
B.1	Sample screen from Experiment 2	392
B.2	Experiment 2 results by item	396
C.1	Sample screen from Experiment 3	400
C.2	Experiment 3 results by item	403
D.1	Sample screen from Experiment 4	407
D.2	Experiment 4 results by item	411
E.1	Sample screen from Experiment 5	415
F.1	Sample screen from Experiment 6	421
G.1	Sample screen from Experiment 7	423
I.1	Sample screen from Experiment 8	438
J.1	Experiment 9 results: Relative duration (leveled data)	448
J.2	Experiment 9 results: Relative duration (all data)	448
K.1	Sample screens from Experiment 10	451
L.1	Sample screens from Experiment 11	453
M.1	Sample screens from Experiment 12	466
N.1	Sample screens from Experiment 13	470

O.1	Sample screens from Experiment 14	472

LIST OF TABLES

3.1	Sample stimuli for Experiment 1	56
3.2	Sample stimuli for Experiment 2	70
3.3	Sample stimuli for Experiment 3	78
3.4	Sample stimuli for Experiment 4	88
4.1	Compatibility ratings for the identity interpretation strategy	133
4.2	Compatibility ratings for the discourse interpretation strategy	144
4.3	Bayes factor interpretation guidelines.	162
4.4	Bayes factors: Hybrid model without error vs. identity and salience models . . .	162
4.5	Bayes factors: Hybrid model with error vs. identity and salience models	164
4.6	Bayes factors: Hybrid model with error vs. hybrid model without error	164
5.1	Percentage deaccenting for r- and l-givenness (Baumann and Riester, 2012) . . .	217
6.1	Sample stimuli for Experiment 8	247
6.2	Sample entailment stimuli for Experiment 9	255
6.3	Sample implicational bridging stimuli for Experiment 9	255
6.4	Sample stimuli for Experiment 10	287
7.1	Sample stimuli for Experiment 11	308
7.2	Experiment 11: Paired comparisons for verb status analysis	317
7.3	Experiment 11: Paired comparisons for accent status analysis	318
7.4	Experiment 12: Model summaries, verb status comparison	334
7.5	Experiment 12: Paired comparisons for verb status analysis	335
7.6	Experiment 12: Model summaries, accent status comparison	335
7.7	Experiment 12: Paired comparisons for accent status analysis	336
7.8	Experiment 13: Model summaries, verb status comparison	336
7.9	Experiment 13: Paired comparisons for verb status analysis	337
7.10	Experiment 13: Model summaries, accent status comparison	337
7.11	Experiment 13: Paired comparisons for accent status analysis	338
H.1	Bayes factors: Salience model with error vs. remaining models	436
H.2	Bayes factors: Salience model without error vs. identity models	436
H.3	Bayes factors: Identity model with error vs. identity model without error	436
I.1	Experiment 8 results	439
I.2	Experiment 8 results, perception items	439
I.3	Experiment 8 results, production items	439
I.4	Entailment items: Mean verb inferability score by item	440
I.5	Implicational bridging items: Mean verb inferability score by item	441
J.1	Speech error trials excluded from phonetic analysis	442

ACKNOWLEDGMENTS

This dissertation, like any other, would have been impossible to complete without a massive degree of academic and social support from the members of my department, the university and the field more generally, and family and friends outside of academia.

I owe by far the biggest thanks to my advisor and committee chair, Ming Xiang. It would be impossible to overstate the patience and generosity Ming has shown as I have developed as a scholar and completed this dissertation. She has devoted an incredible amount of time to me and my projects, beyond anything I could have hoped for when choosing a program and advisor. We have spent countless hours in meetings discussing my work, often sitting in perplexed silence as we try to work out a new experimental paradigm to test a research question or puzzle out what new data is telling us. She has answered emails at all hours of the night as I irresponsibly rush to meet deadlines, and has kept me on track in the program and in my career even when I have faltered. She has approached my project as an opportunity to learn and grow together as researchers rather than a chance to make a student in her own image. It's impossible to imagine a better advisor, and I never could have finished graduate school without her. Thank you, Ming.

I also owe a massive debt of gratitude to the other two members of my committee, Chris Kennedy and Jason Merchant. Ellipsis and deaccenting were nowhere on my radar when I chose to come to Chicago for graduate school, and I am incredibly lucky to have stumbled into a department where I could benefit from the intellect and experience of these two brilliant professors. I thank them both for their time, their valuable critiques of my work, their guidance as I grappled with the (rather intimidating) literature on ellipsis and deaccenting, and their openness to my new methodological approaches and domains of research.

I would also like to thank the rest of the department faculty, especially those who took a particular interest in me or my project, including Diane Brentari, Allyson Ettinger, Itamar Francez, Lenore Grenoble, Jason Riggle, and Alan Yu. I am especially grateful to Alan for his work advising me earlier in my graduate career and for his assistance in designing the

experiments from the second half of the dissertation. Thanks also to my undergraduate instructors, and particularly to Gregory Ward, who first stimulated my interest in linguistic meaning and context.

I am grateful for the contributions of the anonymous reviewers and attendees of the conferences where I presented or will present work from this dissertation: CUNY 30, CUNY 32, ELM 1, ICPHS 19, LSA 92, LSA 93, Meaning in Flux 2, NELS 47, XPrag 7, and XPrag 8. The individual reviews, questions, and discussions resulting from these presentations are too numerous to list, but they have had a substantial impact on the final form of my dissertation. I also owe a debt of gratitude to several researchers who generously offered me their time and valuable comments in meetings outside of these conferences, including Mike Frank, Kyle Johnson, Bob Ladd, Ciyang Qing, and Matt Wagers.

I would like to sincerely thank each of the then-undergraduate students who worked on this project in any way. This includes Kate Christensen, Joe Klafka, Kanan Luce, and Quinn McKissock, as well as any other research assistants in the Language Processing Lab or the department more generally who made contributions.

I would also like to thank the many research participants who completed the experiments presented in this dissertation. As will be obvious to anyone who reads past the acknowledgements section, this work would have been impossible without their attentive participation.

I am grateful to all of the members of the Language Processing Lab, as well as Dan Yurovsky and the members of the Communication and Learning Lab, for enduring many preliminary presentations of this work and providing valuable critique and support.

Graduate school has been a long and, at times, difficult journey, and I am sure it would have been all but impossible without the support, both academic and social, of my fellow graduate students. I would like to thank all of the students from my department and others who offered this support, but I have been particularly impacted by the friendship of Jacob Phillips, Emily Hanink, Ksenia Ershova, Betsy Pillion, and Kat Montemurro, as well as Helena Aparicio, Katie Franich, Julian Grove, Jessica Kantarovich, Asia Pietraszko, Eszter

Rónai, Adam Singerman, and Stephanie Witkowski.

Thanks also to the many students I taught as an instructor or teaching assistant. Teaching at UChicago has been an immensely rewarding experience, and at times has kept me sane when the demands of the research side of grad school felt overwhelming.

Finally, completing this dissertation and graduate school more broadly would not have been possible without the support of my family and friends outside of academia. I would like to thank all of these people, but most of all my mom, who has supported me in every way imaginable throughout this process as well as my life. I suspect she looks forward to continuing to support me in every way except financially.

Whatever value there is to this dissertation I owe to the people above and any who I unfortunately forgot to name. All oversights and errors are my own.

This material is based upon work supported by the National Science Foundation under Grant No. BCS-1827404. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

ABSTRACT

Verb phrase ellipsis and anaphoric deaccenting, although impressionistically quite distinct from one another, present similar theoretical challenges for the interfaces of syntax, semantics, and discourse. For both phenomena, the prior literature has raised questions about the precise division of labor between information from local linguistic antecedents and information from the discourse at large in their licensing and interpretation. The goal of this dissertation is to use novel experimental methodologies to investigate the empirical status of antecedent versus broad discourse information in these constructions, and to use the experimental results to constrain the hypothesis space for the mechanisms underlying their licensing and interpretation.

In the literature on verb phrase ellipsis, there is substantial debate as to whether ellipsis sites stand in a relation of syntactic identity with a linguistic antecedent, or are resolved under semantic identity or along broader discourse lines. This dissertation develops an experimental paradigm to assess verb phrase ellipsis interpretation in “complex discourse contexts,” where the linguistic antecedent and broader discourse context make different information available for inclusion in the interpretation of an ellipsis site. This is followed by the development of a preliminary probabilistic model of verb phrase ellipsis interpretation in context. The results of both investigations agree that verb phrase ellipsis in context is best understood as hybridizing interpretation with respect to the linguistic antecedent and with respect to the general discourse salience of the competing propositions.

In the case of deaccenting, it is widely accepted that material can be deaccented on the basis of being notionally “given” not only when it is identical to an antecedent, but also when its meaning can be inferred, formally or informally, from the meaning of an antecedent. There is disagreement in the literature on whether to represent the deaccenting of antecedent-nonidentical material in the grammar of deaccenting, or as an extragrammatical operation that “rescues” deaccented tokens that fail to meet a grammatical requirement for identity with an antecedent. The dissertation presents a production and perception investigation

of the prosody of two types of inferable verbs: verbs whose meanings are entailed by an antecedent, and verbs whose meanings are made available by an informal inferencing relation linking them to their antecedents. The results reliably indicate that repeated and inferable material have a different empirical status, supporting a licensing account that generates deaccented identical material in the grammar and accounts for deaccented inferable material through an extragrammatical process.

The results of each investigation indicate that each construction instantiates more complex interactions between linguistic structure and discourse information than have previously been considered at length in the literature. The dissertation separately outlines the grammatical and extragrammatical mechanisms that remain viable as explanations for the behavior of verb phrase ellipsis and anaphoric deaccenting in light of the new data. Following this, the conclusion outlines the prospect for a unified analysis of both constructions, with both holistically considering the discourse at large in the grammar or both augmenting a grammatical identity requirement with extragrammatical consideration of discourse.

CHAPTER 1

INTRODUCTION

A major topic of research in linguistic theory concerns the mapping between linguistic form – morphemes, words, and the sentential structures in which they are used – and the meanings that interlocutors communicate and infer. One domain where this mapping is complicated is that of context-dependent expressions. Here, this term is used to refer to linguistic expressions whose licensing or interpretation depends in some way on the discourse context outside the expression itself. Felicitously using and interpreting these expressions might require reference to the prior linguistic context, the nonlinguistic context in which the utterance was used, or both.

This dissertation focuses on two such context-dependent expressions, verb phrase ellipsis and deaccenting. On the surface, these two constructions may appear highly distinct, and it may not be apparent why they should be treated together in one dissertation. Elliptical constructions are characterized by the (apparent) absence of linguistic form, and an obvious question is how listeners readily interpret ellipsis sites in the absence of overt cues. Deaccenting, by contrast, concerns the prosodic encoding of constituents to mark their status as informative or uninformative in the discourse, while the interpretation of the target constituents themselves is largely not in question.

Despite these differences, there are empirical and analytical uniformities between ellipsis and deaccenting that make them worth treating together. Empirically, both constructions index redundancy of a particular constituent with respect to the prior discourse. In general, a constituent can only be elided when its meaning is apparent from the prior discourse, although the precise conditions under which ellipsis can occur remain a topic of heated debate. Deaccenting similarly requires that the constituent’s meaning be “given” in the discourse, but its distribution appears to be somewhat less constrained than that of ellipsis. Analytically, there have been many parallel proposals for the requirements for identity between an ellipsis site and its antecedent and the conditions on deaccenting. The similarities are

substantial enough that some researchers have sought to provide a unified analysis of ellipsis and deaccenting according to which ellipsis is simply a more constrained, “radical” form of deaccenting, but that the two share the same fundamental constraints (e.g., Rooth, 1992a; Tancredi, 1992; Chomsky and Lasnik, 1993).

A major question for both constructions is the precise division of labor between local linguistic antecedents and the broader discourse context, both linguistic and nonlinguistic, in their licensing and interpretation. On the one hand, both types of expression have been argued to depend on relatively strict identity of form with a local linguistic antecedent. On the other hand, it has also been proposed that each can be sensitive to information from the discourse context at large, including prior linguistic information outside a narrow antecedent as well as information from the nonlinguistic context. This dissertation focuses on developing novel experimental paradigms to further investigate the respective empirical contributions of linguistic antecedents and broader contexts to these constructions, and on interpreting the empirical findings in light of the relationship between each construction’s grammar and broader interpretive strategies such as accommodation.

Verb phrase ellipsis (VPE) has been the subject of intensive research over the past several decades. In this construction, a clause is realized without canonically expected material following an auxiliary verb, as in (1).

- (1) Lauren can play the guitar, and Mike can, too.
(Merchant, 2019)

In this sentence, the second clause (barring *too*) ends on the auxiliary verb *can*, despite the canonical expectation that a complement would be required to form a licit structure. Despite this, the intended interpretation of the sentence is clear; the second clause means that Mike can play the guitar, too. Thus, on the surface, elliptical constructions like VPE challenge a fundamental assumption of linguistic theory, namely that there is a principled connection between form and meaning. In such cases, listeners readily infer meaning in the absence of apparent form; this disconnect has inspired much research on the mechanisms by

which listeners interpret elliptical constructions.

Chapter 2 discusses the background literature on VPE, focusing on the relationship(s) relating the ellipsis site to the broader discourse context and possible mechanisms of VPE interpretation. Two classes of account have received much consideration in this literature. According to the first type of account, often called “syntactic identity” accounts, VPE requires a syntactically compatible antecedent verb phrase in the linguistic context, and ellipsis sites should be interpreted as though the antecedent VP appeared *in situ* in the ellipsis site (e.g., Wasow, 1972; Sag, 1976; Williams, 1977; Fiengo and May, 1994). Importantly, this class generally rules out formal “mismatches” with the antecedent, and may predict that information from the broader context should be ignored in interpretation. The second class of account, “semantic identity” accounts, holds that VPE simply requires that an antecedent meaning be available, potentially in the linguistic discourse or the discourse at large (e.g., Dalrymple et al., 1991; Hardt, 1993; Miller and Pullum, 2013). These accounts generally allow for some formal mismatches between the ellipsis clause and the antecedent clause, and also more readily allow for an interpretive mechanism that holistically considers information from the entire discourse.

A related thread in the literature is the question of whether VPE requires a linguistic antecedent in order to be used felicitously and interpreted readily. Early work on VPE largely held that VPE requires a linguistic antecedent, and that apparent uses of VPE in the absence of one (here termed “exophoric” or “antecedentless” VPE) were largely idiomatic (Hankamer and Sag, 1976; Hankamer, 1978; Pullum, 2000). However, more recent work has confirmed that VPE (or VPE-like constructions) can be used productively in the absence of a linguistic antecedent VP, although the distribution of such uses is limited (Schachter, 1977; Merchant, 2004; Miller and Pullum, 2013). This observation potentially has important consequences for the debate between syntactic and semantic identity accounts of VPE; semantic identity, at first glance, is more equipped to deal with ellipsis sites that receive their interpretations from the nonlinguistic context, although there have been proposals for accounting for this

while maintaining syntactic identity.

Chapter 2 highlights a potentially informative empirical gap in the prior research that has been conducted on VPE. The bulk of the literature deals with VPE sites preceded by an exclusively linguistic discourse. By contrast, research on antecedentless ellipsis has largely focused on discourses in which there is no viable linguistic antecedent, as the main question was whether the nonlinguistic context could provide the ellipsis site’s interpretation. As such, prior research has largely not considered VPE in what are referred to here as “complex discourse contexts” – that is, contexts in which there is a viable linguistic antecedent VP, but also salient information in the broader context, including the nonlinguistic context, that may be considered during interpretation. Investigation of such contexts may lead to important insights about the nature of the identity relation for VPE, since it can indicate whether broader contextual information is considered as a rule or generally ignored when there is a viable linguistic antecedent.

Chapter 3 focuses on the experimental investigation of VPE interpretation in complex discourse contexts. The chapter features a novel experimental paradigm in which a linguistic antecedent and a broader discourse context both vary in the degree to which they make number information modifying a particular proposition available. Participant interactions indicate whether they considered a number-modified or -unmodified interpretation of a particular target sentence, VPE or otherwise.

The chapter consists of four experiments yielding the following findings. (i) VPE interpretation in complex discourse contexts does consider information from the broader context when there is no linguistic antecedent; the broader context is considered, albeit more weakly, in the presence of a linguistic antecedent, especially when the broader context makes information highly salient that was not present in the linguistic antecedent. (ii) This critical nonidentity effect for interpretation in the context of a viable linguistic antecedent is at least partially unique to elliptical constructions, and is observed less strongly for fully realized sentences. (iii) There is not a straightforward mapping between the availability of

competing propositions in the discourse at large and their availability under ellipsis; rather, propositions that are compatible with a linguistic antecedent are considered more strongly than is supported by their general discourse availability. (iv) As predicted in the literature, the interpretation of the verbal anaphor *do that* is more sensitive to broader contextual information than is VPE interpretation.

Based on these findings, Chapter 3 rules out two simple accounts of context-situated VPE interpretation: a simple syntactic identity account that ignores broader contextual information when there is a viable linguistic antecedent VP, and a simple semantic or discourse-driven account where VPE interpretations are considered in proportion to their availability in the discourse. It is argued that the interpretive mechanism is more complicated, and two possibilities are highlighted. According to the first, the grammar of VPE mandates syntactic identity with a linguistic antecedent, but alternative antecedents considering information from the broader context can be constructed under certain conditions and used to arrive at a secondary interpretation. The second is a semantic or discourse account with a more articulated set of heuristics for identifying the antecedent meaning such that consideration of the linguistic antecedent is preferred beyond its mere contribution to discourse salience.

Chapter 4 aims to confirm the conclusions of Chapter 3 using a probabilistic modeling methodology. First, three experiments replicate the previous findings using a forced-choice methodology, allowing for interpretation of participant behavior in terms of their probability of accessing particular propositions in the discourse at large and under ellipsis. Following this, six probabilistic models are considered that transform the empirical “prior” probabilities of accessing number-modified or -unmodified propositions in the discourse at large to predicted “posterior” probabilities of accessing the same under ellipsis.

Three classes of model are considered. In identity models, interpretation proceeds only according to the content of the linguistic antecedent. In discourse models, interpretation proceeds according to the empirical availability of the competing interpretations in the context. In hybrid models, the predictions of these strategies are combined in proportion to

a free parameter. In addition, each class of model had one instantiation that provided for “noise” in interpretation, schematized as participants choosing an interpretation at random from a list of possibilities, and one instantiation that did not include random choice.

Comparison of the maximum likelihood instantiation of each model (it’s “peak” performance as a means of predicting the empirical interpretation data) as well as comparison abstracting over possible parameter values agreed that a hybrid model best captured the facts of context-situated VPE interpretation. This finding is interpreted as supporting either of the two interpretive mechanisms proposed after the Chapter 3 experiments, and the chapter closes with a more nuanced discussion of the interface between the models’ specifications and the actual behavior of the interpretive mechanism.

The second half of the dissertation focuses on a phenomenon that will be referred to as “anaphoric deaccenting,” or simply “deaccenting.” This term refers to a prosodic phenomenon in English where certain constituents can receive de-emphasized realizations when they are given in the discourse. In other words, constituents that would normally receive, for instance, a high pitch accent according to the default prosodic rules of English instead receive a low pitch accent or no pitch accent at all when the speaker is obligated to or chooses to mark them as discourse-given.

An example of canonical deaccenting is shown in (2). In examples in this dissertation, small caps are used to indicate constituents that are critically accented – that is, that have a high pitch accent, as a result of, for instance, being in nuclear position in their clause or being contrastively focused. By contrast, italics are used to denote constituents that are deaccented – these are contentful constituents that could potentially host a high pitch accent, but which are realized without accent in the sentence by virtue of their relationship to the broader discourse. This annotation is generally only employed in clauses that contain a deaccented constituent, i.e., not in antecedent clauses.

(2) Bill saw Sue. Then, *MARY* *saw* *JOHN*.

(Tancredi, 1992)

In a canonical realization of (2), *saw* in the second sentence is generally realized without accent. This is despite the fact that *saw* in this position is generally capable of hosting a pitch accent; for instance, *saw* in the first sentence is likely to be realized with a prenuclear pitch accent. The intuition behind this phenomenon is that the meaning of *saw* is old or accessible in the discourse by virtue of the same word having been used in the previous sentence; this licenses deaccenting of the word, perhaps as a means of guiding the listener toward new information in the second sentence that will require additional processing resources to comprehend.

Chapter 5 reviews the background literature on deaccenting, with a focus on the conditions under which constituents can be deaccented. Several complications have prevented the development of a straightforward account of this mechanism. The main one considered here is the claim that constituents can be deaccented not only when they are identical to an antecedent constituent, but also when their meanings are made available by a nonidentical antecedent constituent. An example is shown in (3).

- (3) Yesterday, Sam killed Bill. Today, JOHN *died*.
(Tancredi, 1992)

In this example, *died* is deaccented even though there is not a second instance of *died* in the prior linguistic discourse. Intuitively, this is possible because the first sentence, *Sam killed Bill*, entails that *Bill died*. This makes the meaning of *died* available in the discourse, despite the fact that the expression itself was not used, and deaccenting of a subsequent instance is licensed. Constituents like *died* in the second sentence will be referred to as “inferable” in this dissertation. However, it should be noted here that the label of “inferable” will be applied to constituents whose meanings are not always entailed by an antecedent; “inferable” meanings could also be those that are not logically entailed by the antecedent, but that a particular speaker would nevertheless conclude, or those that a listener would accept via inference as having been given in the discourse upon noting that the meaning was presented as given through deaccenting.

While virtually all researchers of deaccenting mark such examples as acceptable, there are two distinct classes of model of deaccenting licensing that account for these cases in different ways. One class of account, here called the “grammatical” account, makes direct reference to inferencing relations in the grammar of deaccenting (Ladd, 1980; Selkirk, 1984; Rochemont, 1986; van Deemter, 1994, 1999; Schwarzschild, 1999; Büring, 2016, i.a.). The availability of a constituent’s meaning in the linguistic or broader discourse is the very thing that makes it deaccentable, so deaccenting is automatically ruled in for constituents whose meanings are entailed in the prior discourse (or made available by other more informal inferencing relations, although the details underlying these are not always spelled out).

The second class of model is referred to as the “accommodation” model. According to this account, the grammar of deaccenting licensing actually requires that a deaccenting target have an identical correlate in the linguistic antecedent. Thus, deaccenting of constituents that do not have identical correlates is strictly ungrammatical. However, deaccented nonidentical material can be salvaged and judged as acceptable through the construction of a covert alternative antecedent against which to calculate identity (Tancredi, 1992; Fox, 2000). This operation is triggered exactly by the presence of ungrammatically deaccented material, and is presumably licit as long as the meaning of the accommodated antecedent is reasonable in the discourse context.

While sentences with deaccented antecedent-nonidentical material are commonly treated as a core piece of data to be accounted for by the deaccenting licensing mechanism, these examples actually receive relatively little empirical attention in the literature, and the judgments provided are mostly introspective. There has been little systematic empirical investigation of these sentences, particularly with respect to the acceptability of deaccenting inferable material relative to canonical deaccenting under identity with an antecedent constituent. One exception to this trend is a study of the prosodic realization of nouns under various discourse statuses that was performed by Chodroff and Cole (2019) roughly contemporaneously with the research presented in this dissertation; the results of that study are

problematic for the empirical generalization noted in the literature and are highly compatible with the data presented in this dissertation.

Chapters 6 and 7 focus on filling this empirical gap by systematically investigating the relative status of accentuation and deaccentuation of discourse-new, given, and inferable constituents. The goal is to establish the empirical status of accented and deaccented inferable constituents relative to more canonical repeated and discourse-new constituents, and to interpret these findings in light of the theory of deaccenting licensing. The experimental paradigm focuses on sentences of the form *subject verb object and subject verb object*, or *SVO and SVO*. The critical focus is on the verb in the second clause, which through manipulation of the first-clause verb can be made discourse-new, discourse-given (identical to the first-clause verb), or inferable. Two types of inferencing relation are investigated. In some items, the meaning of the second verb is entailed by that of the first verb (modulo type shifting), as in the pair *hugged/embraced*. In the other items, the relationship is more informal. The second verb is not necessarily entailed by the first, but upon hearing the second verb, participants might agree that the meaning of the second verb is made reasonably available through the use of the first, as in the pair *charmed/seduced*.

Chapter 6 focuses on the production of sentences in this experimental paradigm. First, a norming study establishes a set of experimental stimuli in which the inferencing relation between first- and second-clause verbs in the “inferable” condition makes the meaning of the second verb highly inferable, whereas the meanings of second-clause discourse-new verbs are verified to be non-inferable. The principal experiment of the chapter is a production study in which participants read highly controlled *SVO and SVO* sentences with the critical manipulation. Prosodic analysis of intensity, f_0 , and duration on the second-clause vocalic nuclei reveals that participants reliably accented both new and inferable verbs and deaccented repeated verbs. A follow-up study eliciting judgments of emphasis on second-clause words from naive listeners confirmed these findings. The experimental results are noted to challenge the empirical generalization claimed in the literature that deaccenting of nonidentical, but

inferable, constituents is straightforwardly acceptable. However, the results are also noted to be more compatible with the accommodation account than the grammatical account, since the marking of such examples as ungrammatical in the accommodation account leaves room for an acceptability gap between these cases and canonical cases of deaccenting under nonidentity.

Chapter 7 continues the investigation by focusing on perception. In each of the four experiments in this chapter, recorded stimuli were manipulated so that a canonically realized accented or deaccented verb could be used in a context making it new, inferable, or repeated. The first experiment elicits judgments of these sentences out-of-the-blue, and the results largely parallel those of the production study, with new and inferable verbs patterning together with a preference for accentuation, while repeated verbs show a preference for deaccentuation.

The subsequent experiments focus on the role of the broader discourse context in modulating the acceptability of deaccented inferable material. Two experiments investigate the role of a prior context that moderately supports a “situationally identical” reading of the antecedent and target verb. These contexts eroded the preference for accentuation on inferable verbs, and participants were indifferent between accentuation and deaccentuation. The final experiment added the presupposition trigger *too* to the end of the experimental stimuli; this resulted in a preference for deaccentuation of inferable verbs. Thus, the acceptability of deaccented inferable material sits on a cline and varies with the degree of support in the broader context for identical readings of the antecedent and target. These findings are argued to be more compatible with the accommodation account of deaccenting under nonidentity, although it is noted that a grammatical account that assumes that calculating inference relations linking antecedents and nonidentical deaccenting targets is also plausible.

Finally, Chapter 8 summarizes the findings of the previous chapters, considers similarities and differences in the remaining plausible analyses of verb phrase ellipsis and deaccenting, discusses possible implications for the broader theory of anaphora and context dependence,

and highlights some possible future directions of research. Both VPE and deaccenting are noted to invoke complex interactions between their linguistic antecedents, when present, and the broader discourse context. The discussion touches a classic division posited in the literature between *deep* and *surface* anaphoric constructions (Hankamer and Sag, 1976; Sag and Hankamer, 1984). Deep anaphors can freely receive their interpretations from the prior discourse at large, while surface anaphors require identity with a linguistic antecedent.

The analyses of VPE and deaccenting that were not ruled out by the experimental results of the dissertation are noted to be “deep” or “surface” in character. The complex discourse-driven VPE interpretation mechanism and the grammatical account of deaccenting licensing are “deep” in the sense that they holistically consider information from the discourse context at large directly in the grammatical mechanism. By contrast, the “augmented” identity account of VPE and the accommodation account of deaccenting are “surface” in that they posit that the grammar of the respective operations depends on strict identity with a linguistic antecedent, with any acceptable deviation from the antecedent due to an extragrammatical operation such as accommodation.

It is suggested that future research that can definitively settle on one mechanism or the other for either construction might contribute to the broader understanding of deep and surface anaphors. On the one hand, if the mechanisms are really “deep,” it might suggest that the deep-surface divide is actually illusory; all anaphoric and context-dependent constructions are actually deep, but the heuristics guiding the resolution of certain constructions, such as VPE and deaccenting, heavily favor the consideration of information from the linguistic antecedent and make them appear surface. On the other hand, if the mechanisms really are “surface,” it suggests that the surface-deep distinction is real and divides constructions that rely grammatically on identity with a linguistic antecedent, like VPE and deaccenting, from truly deep constructions like verbal anaphora and pronouns that can freely receive their interpretation from the broader context. The conclusion highlights consideration of timing data gathered through online methodologies, the application of identical experimental

methodologies to multiple different types of construction, and the investigation of yet more articulated discourse contexts as possible future directions of research that may inform this question.

CHAPTER 2

BACKGROUND: IDENTITY AND NONIDENTITY IN VERB PHRASE ELLIPSIS

The first half of this dissertation focuses on the interaction of linguistic antecedents and information from broader discourse contexts in the interpretation of verb phrase ellipsis. Ellipsis has received considerable attention in the linguistic literature over the past several decades. By its very nature, ellipsis challenges the traditional linguistic assumption that there is a principled link between linguistic form and understood meaning – listeners infer meaning from ellipsis sites, which in form appear to be contentless. The fact that ellipsis sites can be interpreted suggests the possibility of meaning without form; to “rescue” this form-meaning mapping requires positing form or structure that is not audible (Merchant, 2019).

Three major questions have emerged in the ellipsis literature: the structure question, the recoverability/identity question, and the licensing question (Merchant, 2019). The structure question investigates whether ellipsis sites contain unpronounced syntactic structure. The recoverability or identity question asks whether the relationship between ellipsis sites and their antecedents is one of syntactic or semantic identity. The licensing question investigates the structural and discursive conditions under which ellipsis can felicitously be used.

This dissertation focuses on the identity question, although all three questions are intertwined and structure and licensing are also discussed below. Two major approaches to the identity question have emerged in the literature. The first approach holds that ellipsis sites are obligatorily related to a linguistic antecedent by a relation of syntactic identity, modulo differences in minor domains like inflectional morphology (Wasow, 1972; Sag, 1976; Williams, 1977; Fiengo and May, 1994, i.a.). The second holds that ellipsis sites require a semantic object as an antecedent, either the meaning of an antecedent verb phrase or, under some accounts, a meaning that is available in the discourse at large (Dalrymple et al., 1991;

Hardt, 1993; Merchant, 2001, 2018, i.a.).

The main piece of evidence that has been used to argue for one approach or the other is sentences with formal mismatch between the ellipsis clause and a linguistic antecedent. For instance, much attention has been paid to sentences where the clauses differ in voice, such as a sentence with a passive first clause and an ellipsis remnant that forces an active interpretation. Judgments on such examples vary considerably, from fully acceptable to fully unacceptable. When formal mismatches between semantically equivalent constructions lead to unacceptability, it is argued to support a requirement for syntactic identity; when formal mismatches are permissible, it is considered to support a semantic identity account.

One extreme type of mismatch that has received relatively little attention until recently is the case of exophoric or antecedentless ellipsis. Under certain discursive conditions, verb phrase ellipsis can be used in the absence of a viable linguistic antecedent VP, in which case it is interpreted with respect to the information that is salient in the nonlinguistic context. The possibility of interpreting VPE in the absence of prior syntactic structure has been argued to constitute evidence for a semantic identity or discourse-driven answer to the identity question, although other accounts have been proposed that allow for the maintenance of a syntactic identity grammar for VPE.

This dissertation targets an empirical gap in the prior literature on VPE, inspired by the literature on exophoric ellipsis, that may inform the identity question. Prior work on VPE has generally focused on exclusively linguistic contexts, and among these, generally single sentences. The research on exophoric ellipsis likewise focuses more or less exclusively on cases of VPE with no viable linguistic antecedent, concentrating on the question of whether VPE can indeed be felicitously used based only on support from the nonlinguistic context.

This raises the question of how ellipsis is interpreted when there is both a viable linguistic antecedent VP as well as highly salient information in the broader context that can contribute to interpretation. If information from the broader context is ignored in such contexts, it suggests that VPE is interpreted under syntactic identity. If information from the broader

context is considered even in the presence of a viable linguistic antecedent, it suggests that a semantic or discourse process plays a role in interpretation.

The remainder of this chapter reviews the relevant background on verb phrase ellipsis, focusing on the evidence for syntactic and semantic accounts of identity and the development of the evidence for and analysis of exophoric ellipsis. Following that, the next chapter experimentally investigates the interpretation of verb phrase ellipsis in complex discourse contexts, and the chapter after focuses on the development of a probabilistic model of context-situated verb phrase ellipsis interpretation.

2.1 Evidence for syntactic identity

Syntactic identity accounts hold that the relationship between ellipsis sites and their linguistic antecedents is one of identity of form. In essence, this means that the ellipsis site contains unpronounced syntactic structure at some point in the derivation that is identical to that in the antecedent VP, and that the ellipsis site is interpreted as though the exact antecedent VP were situated in the ellipsis site, modulo minor differences such as inflectional morphology. Syntactic identity accounts can posit that the ellipsis site contains null elements that are replaced by copies of antecedent structures during interpretation (Wasow, 1972; Williams, 1977; Fiengo and May, 1994), or that some operation causes the nonpronunciation of otherwise present syntactic structure (Ross, 1969; Sag, 1976; Merchant, 2001).

The main evidence for a syntactic answer to the recoverability question for VPE comes from elliptical sentences whose acceptability or distribution of available interpretations is affected by syntactic mismatches between the ellipsis site and its linguistic antecedent, where there is no corresponding semantic mismatch that can be appealed to. One of the most cited types of evidence in the literature on VPE is sentences where the antecedent clause and the elliptical clause differ in voice, as in (4).

- (4) #This problem was looked into by John, and Bob did too.
(Kehler, 2000)

Many of these voice mismatch examples are commonly judged to be fully unacceptable, although intuitions regarding individual sentences appear to depend on a variety of factors, and some instances are marked as having middling or strong acceptability. This spectrum of acceptability forms a key part of the debate between a syntactic and a semantic account of recoverability, and will be discussed in more detail below.

The fundamental insight behind examples like (4) is that the mismatch in voice appears to be directly responsible for the degradation in acceptability, as is apparent when compared to voice-matched examples like (5).

- (5) John looked into this problem, and Bob did too.

This effect points to a role of syntax in determining what types of linguistic structures can serve as antecedents to VP ellipsis sites. Active and passive counterparts like *x looked into this problem* and *this problem was looked into by x* are mutually entailing and are generally treated as semantically identical to one another (Merchant, 2013). Thus, it is difficult to account for the unacceptability of (4) relative to (5) on an account where ellipsis relies on a semantically isomorphic structure in the antecedent; since the antecedent and putative elided VPs are identical in meaning, a simple iteration of the semantic analysis predicts that examples like (4) should be acceptable. By contrast, the two VPs clearly have different syntax, so an analysis that requires syntactic isomorphism between ellipsis and antecedent correctly rules such examples out.

A similar type of evidence comes from sentences that are acceptable, but which do not exhibit interpretations that are predicted by a semantic identity analysis. Such cases often focus on binding and co-indexation among proper names and pronouns; two examples are shown in (6) and (7).

(6) *The lawyer defended Bill_i, and he_i did too. [defend Bill_i]

(Lappin, 1992)

(7) *John_i blamed himself_i, and Bill did too. [blame John_i/him_i]

(Kitagawa, 1991)

Both of these examples suggest that the ellipsis recoverability relation is sensitive to syntax rather than semantics. In each case, the indicated interpretation for the ellipsis site is unavailable despite the fact that the inferred object is co-referential with the object in the antecedent clause. This is explained straightforwardly on a syntactic identity account, since interpreting the ellipsis site with the exact antecedent VP *in situ* would lead to a syntactic violation. In (6), a proper name would be c-commanded by a co-indexed pronoun, resulting in a Condition C violation; in (7), the second-clause reflexive would be co-indexed with a nonlocal antecedent, resulting in a Condition A violation.

Beyond the evidence from syntactic mismatches and binding indicating a sensitivity to syntax in ellipsis interpretation, there is considerable evidence that ellipsis sites contain covert, articulated syntactic structure (i.e., the structure question). On an intuitive level, this may seem to suggest that the ellipsis identity relation is also syntactic. However, it is important to note that the presence of structure in the ellipsis site does not necessarily entail that ellipsis sites are resolved under syntactic identity, and a number of proposals allow for both unpronounced structure and semantic identity (Sag and Hankamer, 1984; Merchant, 2001; van Craenenbroeck, 2010; Aelbrecht, 2010). On the other hand, Frazier (2019) suggests that from a processing perspective, the presence of structure in ellipsis sites is likely indicative that ellipsis sites are interpreted by copying syntactic structure from the antecedent.

Much of the evidence for syntactic structure in ellipsis sites considers forms of ellipsis other than VPE, such as sluicing and fragment answers. However, some experimental investigations of this question have considered evidence from VPE. For example, Xiang et al. (2019) found evidence of syntactic priming from ellipsis sites on subsequent syntactic pro-

duction. They primed participants with sentences like those found in (8) and (9), then had the participants describe a picture that could be described with a prepositional dative or double object construction.

(8) First Ralph sang a song to Sheila, and then...

- a. Marcus sang one to her. [nonelliptical]
- b. Marcus did. [VP ellipsis]
- c. Marcus groaned. [neutral control]

(9) First Ralph sang Sheila a song, and then...

- a. Marcus sang her one. [nonelliptical]
- b. Marcus did. [VP ellipsis]
- c. Marcus groaned. [neutral control]

Importantly, under an account where interpreters construct syntactic structure in the ellipsis site under identity with a linguistic antecedent, both the nonelliptical and VP ellipsis primes contain two instances of either a prepositional dative (8) or a double object construction (9). By contrast, the neutral control sentences contain only one such construction. The production results showed that participants were primed to favor the prepositional dative or double object construction to similar degrees with nonelliptical and VP ellipsis prime sentences, but that priming was not observed in the neutral control conditions. The results suggest that priming in this paradigm required a construction like the nonelliptical sentences, with either a second or recent instance of the construction of interest, but that the VP ellipsis site in the elliptical sentences “counted” as such an instance, suggesting the presence of syntactic structure in the ellipsis site.

Other experimental studies have investigated whether lexical access occurs separately inside ellipsis site in addition to upon first exposure in a nonelliptical antecedent. Shapiro and Hestvik (1995) had participants listen to conjoined VPE sentences such as (10) and

perform a lexical decision task at various points, denoted in the example by numbered positions.

- (10) The policeman defended himself, and the fire[1]man did [2] too, according to someone [3] who was there.

At position [2], the location of the ellipsis site, Shapiro and Hestvik recorded lower reaction times in lexical decision for words that were semantically related to the first-clause subject (e.g., *robber*) compared to unrelated words (e.g., *roller*). They take this to indicate that the first-clause subject is reaccessed at the time position [2] is being processed, suggesting syntactic reconstruction in the ellipsis site. Importantly, the effect cannot be explained in terms of continuing activation from the initial processing of the first-clause subject, since no reaction time difference was observed for related versus unrelated probes at position [1].

In a similar vein, Snider and Runner (2014) used a visual world eye tracking paradigm to assess the activation of semantic and phonological neighbors in the interpretation of sentences like those shown in (11).

- (11) The security guard opened the lock, and the night watchman...
- a. did, too. [VP ellipsis]
 - b. did it, too. [VP anaphora]
 - c. dropped it. [anaphora]
 - d. slept. [intransitive control]

In the VPE condition, participants looked to both semantic (*key*) and phonological (*log*) competitors for the first-clause object *lock* significantly more than in the intransitive control condition. Both of these effects are well established correlates of lexical access (Allopenna et al., 1998; Yee and Sedivy, 2006), and the lack of such an effect in the intransitive condition indicates that consideration of these competitors was not simply due to carry-over activation of *lock* from the first clause. Snider and Runner conclude from this that *lock* is accessed a second time during the interpretation of the ellipsis site.

Beyond these investigations of the online processing of VPE, there is substantial introspective evidence for syntactic structure in ellipsis sites. Some of these findings have been noted for VPE sites, whereas other arguments use data from other types of ellipsis, like sluicing. While in principle the answer to the structure question might be different for VPE versus other types of ellipsis, there have been efforts to present a uniform analysis for the purposes of theoretical economy (Merchant, 2013, 2019). Thus, such evidence is worth mentioning here.

Some commonly cited pieces of evidence of syntactic structure in ellipsis sites are the sensitivity of elliptical sentences to islands, the P-stranding generalization, and case-matching effects. First, it is well established that movement in ellipsis sentences such as (12) is generally subject to the same island constraints as in fully realized sentences (Haik, 1987).

- (12) *Abby wants to hire someone who speaks a Balkan language, but I don't remember which Balkan language Ben does. [want to hire someone who speaks *t*]
(Merchant, 2004)

To the extent that island constraints must be given a syntactic analysis, this finding suggests that ellipsis sites must contain the relevant syntactic structure to block the extraction.

Second, Merchant (2001) presents the P-stranding generalization, which notes that exactly those languages that allow P stranding in general, like English in (13), will allow stranded prepositional remnants outside of sluicing or fragment answer ellipsis sites. By contrast, languages that do not generally allow P-stranding, like Greek in (14), will also disallow prepositions stranded outside ellipsis sites.

- (13) a. Peter was talking with someone, but I don't know (with) who(m).
b. Who was he talking with?

- (14) a. I Anna milise me kapjon, alla dhe ksero *(me) pjon.
the Anna talked with someone but not I.know with who
b. *Pjon milise me?
who talked.3S with

These facts are readily explained by an account that posits syntax in the ellipsis site, since whatever syntactic constraint governs the distribution of P-stranding in these languages in general is simply assumed to operate over elided structure, as well.

Finally, in languages that mark case, remnants show effects that appear to derive from elided structure in both sluicing (Ross, 1969) and fragment answers (Merchant, 2004). For instance, while Ross’s example in (15) shows accusative case on the remnant, which might otherwise be expected for a notional object, (16) shows that the remnant must have dative case exactly when the elided verb idiosyncratically assigns it.

- (15) Er will jemanden loben, aber sie wissen nicht, { *wer / wen /
 he wants someone.ACC praise but they know not who.NOM who.ACC
 *wem }.
 who.DAT

- (16) Er will jemandem schmeicheln, aber sie wissen nicht, { *wer / *wen /
 he wants someone.DAT flatter but they know not who.NOM who.ACC
 wem }.
 who.DAT

Again, the analysis here is straightforward if the ellipsis site is assumed to contain syntactic structure that assigns case according to exactly the same principles as in nonelliptical constructions.

Merchant (2019) and Frazier (2019) review considerable additional evidence for syntactic structure inside ellipsis sites. For the purposes of this background, it is sufficient to say that the evidence for structure preponderates, and suggests that ellipsis sites are also resolved under identity, to the extent that such a link is valid (e.g., Frazier’s (2019) processing claim).

2.2 Evidence for semantic identity

As mentioned above, the main alternative to syntactic identity accounts of the resolution question are semantic identity accounts. According to these models, the relationship between an ellipsis site and its antecedent (which, in some accounts, need not be a linguistic object *per se*) consists in identity of meaning. In essence, if the antecedent and ellipsis are expressed

in semantically equivalent terms, then the ellipsis should be licit and interpretable, even if there is a formal (syntactic) mismatch between the two. The details of the analysis vary by account. Several take the tack that a VP ellipsis site can be treated essentially as a null proform which is resolved using mechanisms similar to those employed for other anaphors, including pronouns (Schachter, 1977; Chao, 1987; Hardt, 1992; Kehler, 1993a; Lobeck, 1999). Another approach is to define a formal meaning relation between the ellipsis site and its linguistic antecedent, such as Merchant’s (2001, 2018) proposal that ellipsis is licensed when the inferred elided expression and its antecedent are mutually entailing, modulo type shifting.

The most important evidence in favor of such accounts is a collection of examples with formal antecedent-ellipsis mismatches that are nevertheless marked as acceptable in the literature. The label *vehicle change* refers to a class of relatively minor formal mismatches between ellipsis sites and their antecedents that do not substantially degrade the acceptability of the corresponding sentence (Dalrymple et al., 1991; Fiengo and May, 1994). An example of pronoun/name vehicle change in VPE is shown in (17).

- (17) They arrested Alex_i, though he_i thought they wouldn’t.
(Merchant, 2001, 2019)

If the ellipsis site in this sentence were interpreted with the exact syntactic antecedent VP *in situ*, it would be ungrammatical, since the name *Alex* would be c-commanded by the co-referential pronoun *he*. Nevertheless, the sentence is judged acceptable.

Similarly, an example of polarity item vehicle change appears in (18).

- (18) John didn’t see anyone, but Mary did.
(Sag, 1976; Merchant, 2019)

In this case, interpreting the ellipsis site with the antecedent *in situ* would entail placing the negative polarity item *anyone* outside a licensing nonveridical context (Giannakidou, 1998).

Finally, although unacceptable voice mismatch examples were cited above as evidence for a syntactic identity account, some cases, such as (19), have been marked as acceptable or marginal in the literature.

- (19) This problem was to have been looked into, but obviously nobody did.
(Kehler, 1993b)

In essence, the mismatch argument is that it is relatively easy to give a uniform semantic analysis to co-referential names and pronouns, to corresponding positive and negative polarity items, or to active-voice and passive-voice encodings of the same proposition, whereas it is difficult to give a uniform syntactic analysis to any. Proponents of semantic identity take this as evidence that the ellipsis recoverability relation makes reference only to identity of meaning between ellipsis and antecedent, not identity of form. In short, mismatches are grammatical.

Kehler (2019) provides several additional pieces of evidence beyond mismatches that suggest an anaphoric analysis of VPE rather than a syntactic identity one. First, he notes that the distribution of cataphoric VPE (20a, 20c) has long been noted to be parallel to that of cataphoric pronouns (20b, 20d) (Lakoff, 1968; Jackendoff, 1972).

- (20) a. If McCain will, Obama will make a statement criticizing President Putin.
b. If he_i makes a statement criticizing President Putin, Obama_i will make a fool of himself_i.
c. *Obama will, if McCain will make a statement criticizing President Putin.
d. *He_i will make a fool of himself_i if Obama_i makes a statement criticizing President Putin.

Like anaphors, it is possible, although dispreferred, for VPE antecedents to be located quite distantly in the discourse, as in (21). According to an analysis of the Brown corpus conducted by Hardt 1990, around 5 percent of tokens of VPE had an antecedent that was two more sentences earlier in the discourse.

(21) The thought came back, the one nagging at him these past four days. He tried to stifle it. But the words were forming. He knew he couldn't.

(22) shows that like anaphors, VPE can have a split antecedent, where the antecedent constituents corresponding to the ellipsis site's interpretation are separated by intervening material. In this case, the pronoun *them* is interpreted as referring to both Mary and Fred, while the ellipsis site is interpreted as meaning that only one of them can go to their desired destination.

(22) Mary wants to go to Spain and Fred wants to go to Peru, but because of limited resources, only one of them can.

(Webber, 1978)

One final observation, which will be discussed in more detail below and critically informs the empirical investigation in the next two chapters, is that VPE can occasionally occur with no apparent overt VP antecedent, as in (23). This is notably similar to pronouns and other “deep” anaphors, which are known to relatively freely receive their interpretations from any available context, including salient nonlinguistic information, rather than obligatorily being interpreted with reference to a local linguistic antecedent (Hankamer and Sag, 1976).

(23) [John pours another martini for Mary.]

Mary: I really shouldn't.

(Schachter, 1977)

Beyond constituting an additional similarity between VPE and pronouns, such cases of antecedentless ellipsis can be viewed as “radical” mismatches that invite a semantic identity analysis; since there is absolutely no overt linguistic material with respect to which the ellipsis site can be interpreted, it stands to reason that it must be interpreted with respect to the information that is available in the discourse at large. If this mechanism is available in the absence of the antecedent, it would be theoretically economical to propose that it is the same mechanism that underlies interpretation of canonical ellipses with linguistic antecedents.

Kehler notes that all of these properties of VPE are shared with conventional “deep” anaphors like pronouns, to the exclusion of more obviously syntactically sensitive ellipsis phenomena like gapping and stripping. He argues that these similarities may point to an anaphoric analysis of VPE, whereby it is resolved under semantic (and pragmatic) identity with information available in the linguistic and nonlinguistic discourse rather than under syntactic identity with a particular, singular antecedent VP.

2.3 Resolving overgeneration

Mixed acceptability judgments on mismatched antecedent-ellipsis pairs, such as voice mismatch examples, constitute one of the fundamental obstacles to determining the nature of the connection between ellipsis sites and their antecedents. In principle, syntactic accounts of the identity relation undergenerate with respect to the actual data, since they rule out formal antecedent-ellipsis mismatches (besides trivial cases like mismatches in inflectional morphology), but there are attested cases of mismatch that are marked acceptable, or at least more acceptable than clearly ungrammatical sentences. Conversely, semantic accounts overgenerate with respect to the data, since they predict that semantically equivalent expressions should be intersubstitutable in elliptical sentences, yet there are cases where formal mismatch is the only apparent source of degradation in acceptability judgments.

There are two approaches to resolving this tension. The first possibility, discussed here, is to posit that the semantic account is essentially correct, and to explain why some sentences that should be ruled in by the semantic identity relation sound degraded. This generally amounts to locating some other penalty that affects these sentences, perhaps one that is completely separate from ellipsis, that shows their degradation is not due to the formal mismatch itself. The other approach, discussed in the next section, holds that the syntactic account is essentially correct, and that acceptable cases that are ruled out by the grammar of ellipsis itself are “rescued” by some other grammatical or processing consideration.

Kehler (2000, 2002) proposes one account of the former type. Under Kehler’s account,

part of participating in a discourse involves inferring various *coherence relationships* that link clauses into a logical flow. For instance, one clause might describe the cause of the content of another clause, or two clauses might describe similar states of affairs. Kehler’s proposal is that the type of coherence relation linking an antecedent clause and an ellipsis clause has an effect on whether the identity relation is best understood as syntactic or semantic.

If two clauses stand in a *cause-effect* relationship, the proposition of one of the clauses stands in a causal relation with the proposition of the other. An example of a cause-effect relation (Kehler’s “Violated Expectation” subtype) is shown in (24).

(24) Bill was about to be impeached, but he didn’t call his lawyer.

Despite the fact that Bill didn’t actually call his lawyer, the two clauses of (24) stand in a cause-effect relationship because normally one would infer from the fact that someone was about to be impeached that they would call their lawyer. The contrasting type of coherence relation is a *resemblance* relation, where the two constituents are parallel or contrastive in some way. An example of a resemblance relation (Kehler’s “Contrast II” subtype) is shown in (25).

(25) John voted for Clinton, but Mary voted for Dole.

Here, the two clauses are parallel in the sense that they provide the same type of information about both John and Mary, but contrast in the specific property that they attribute to each individual. Kehler claims that cause-effect relations and resemblance relations differ with respect to whether they require both syntactic and semantic or just semantic identity to license ellipsis. Cause-effect relations are defined only according to the propositional meanings of their component clauses. As such, Kehler suggests that semantic identity alone is sufficient to license ellipsis, and syntactic mismatches are permitted in this type of relation, such as the voice mismatch in (26)

(26) This problem was to have been looked into, but obviously nobody did.

By contrast, establishing a resemblance relation involves accessing sub-clausal content in order to evaluate the extent to which the propositions of the clauses resemble or contrast with one another. Because sub-clausal information is at issue in these relations, Kehler claims that syntactic identity is required to license ellipsis, even if the semantic content of the constituents is apparently sufficient to resolve the ellipsis. Mismatches like the voice mismatch in (27) are ruled out.

(27) This problem was looked into by John, and Bob did too.

With this approach, Kehler aims to cut across the mismatch data on a dimension that is orthogonal to ellipsis itself, with independent facts about how listeners establish the links of a coherent discourse determining whether mismatches are tolerated or ruled out.

Kertz (2008, 2013) targets a different dimension of the mismatch data, suggesting that it is canonical versus noncanonical encoding of information structure that divides acceptable mismatches from unacceptable ones. Like Kehler, she proposes that voice mismatch sentences can be divided into two categories, but rather than splitting them based on the type of coherence relation linking the clauses, she targets the focus structure of the sentence.

On the one hand are sentences where the elliptical clause contains a contrastively focused auxiliary, such as (28). According to Kertz, these sentences do not violate canonical information structure encoding principles even when there is a formal antecedent-ellipsis mismatch, and they are ruled in.

(28) A lot of this material can be skipped, and often I do_F.

On the other hand are sentences with contrastively focused topics in the ellipsis clause. An example is shown in (29).

(29) #The material was skipped by the instructors and the TAs_F did too.

The crux of the proposal is that sentences with contrastively focused topics often violate independent conventions of information structure encoding. Typically, a contrastive topic

will appear in subject position. The choice to encode the ellipsis clause of (29) with passive voice and the contrastive topic in object position incurs a processing cost, leading to degraded acceptability for the sentence as a whole. Kertz supports this with a self-paced reading study, arguing that readers actually detect the ill-formedness of sentences like (29) at the subject of the second clause, where the information structure violation becomes apparent, even before they could know that the sentence contains ellipsis.

Another approach to reconciling a semantic identity account with variable judgments of mismatch acceptability appeals to the notion of the Question Under Discussion, or QUD (Beaver and Clark, 2008; Roberts, 2012). Miller and Hemforth (2014) examine a different type of mismatch, where VPE constructions are interpreted using a nominal antecedent, as in (30).

- (30) Mubarak’s *survival* is impossible to predict and, even if he does, his plan to make his son his heir apparent is now in serious jeopardy.

Under Miller and Hemforth’s analysis, VPE interpretations are recovered via a semantic relationship to an antecedent. However, the discourse imposes conditions on when the use of a VPE sentence will sound felicitous. They, together with Miller and Pullum (2013), propose that VPE is only felicitous when the discourse context makes a particular contrast set highly salient, and the VPE sentence serves to pick out one member of the set. In the case of “auxiliary-focus” VPE, a set of propositions that contrast in polarity, tense, aspect, or mood must be made highly salient in the discourse, with the VPE utterance picking out one alternative.

The analysis for (30), then, is that the noun *survival* implicitly raises the contrast between two propositions, *Mubarak survives* and *Mubarak doesn’t survive*, to sufficient salience to license VPE. Since VPE interpretation requires only a semantic antecedent on this analysis, and the ellipsis in this case is licensed by the highly salient contrast, the sentence is judged acceptable despite the impressionistically large formal mismatch between an elided VP and an antecedent noun.

Finally, it has been proposed that parsing preferences can play a role in degrading acceptability ratings for mismatched ellipsis sentences that are grammatical. For instance, Kim et al. (2011) propose a grammar that rules in all antecedent-ellipsis mismatches. It is important to note that their proposal actually requires exact syntactic, not semantic, identity between an ellipsis site and its antecedent. However, they propose a highly articulated clause structure that allows for a high degree of flexibility in construing two constituents as semantically identical. While the proposal is underlyingly syntactic, it shares with the models above the notion that mismatches (at least, voice mismatches) are uniformly grammatical, while independent factors lower the acceptability of certain tokens.

Many examples of mismatched ellipses are reported to have middling acceptability in the literature, and Kim et al. ascribe this to the parsing heuristics employed by comprehenders rather than a grammatical constraint against mismatch itself. A key piece of evidence in Kim et al.’s study is the traditional voice mismatch sentences. They propose that two independently motivated parsing heuristics are brought to bear when processing such sentences. The first is MaxElide, which is a preference for eliding larger constituents over smaller ones (Merchant, 2008; Takahashi and Fox, 2005; Frazier and Clifton Jr., 2005). Functionally, Kim et al. interpret MaxElide as a violable preference for ellipsis to subsume the voice head. Critically, the grammar they assume requires ellipsis at V (below voice) when the antecedent and ellipsis clause differ in voice, but allows ellipsis of voice when the clauses match.

The second heuristic, which is not unique to ellipsis, is Canonical Realization (Bever, 1970; Townsend and Bever, 2001). Essentially, this is a preference for *in situ* argument realization that penalizes structures where underlying subjects are not realized as surface subjects. In voice mismatch, this boils down to a blanket penalty for passives compared to actives. Thus, Kim et al. predict a penalty for the use of passive voice in general, and an additional penalty for ellipsis in a clause with a different voice specification than its antecedent, since this forces ellipsis at V rather than voice. They argue that the parser must investigate additional structures when the configurations preferred by MaxElide and

Canonical Realization do not yield a licit structure, and this additional processing lowers the acceptability of sentences.

Kim et al. test their predictions with an acceptability judgment task investigating active-active, passive-passive, active-passive, and passive-active ellipsis sentences, among others. They found that passive-passive match sentences were dispreferred to active-active match sentences, in keeping with the blanket penalty for passive generated by Canonical Realization. Furthermore, active-passive and passive-active mismatches were both rated worse than passive-passive matches, reflecting the prediction that MaxElide penalizes elliptical sentences where the voice head is left outside the ellipsis site.

2.4 Resolving undergeneration

Just as the accounts discussed above aim to explain the variable acceptability judgments for ellipsis mismatch data by constraining the predictions of a grammar that rules all mismatches in, there have been attempts to do the same by “rescuing” certain examples ruled ungrammatical by a grammar that enforces strict identity between ellipsis sites and their antecedents.

One such approach is presented by Arregui et al. (2006) as the *VP recycling hypothesis*. According to this hypothesis, when a listener encounters a VPE site that does not co-occur with a viable antecedent, the listener can construct a new antecedent using the material “at hand” – that is, other linguistic material available in the context that could help support a reasonable interpretation.

Importantly, they propose that the degree of processing difficulty that is associated with reconstructing an antecedent VP is correlated with the level of acceptability the listener will ascribe to the sentence. If the antecedent was relatively easy to reconstruct – meaning it took relatively few transformations to change some available linguistic material into a viable antecedent – then the listener can have the impression that the sentence is relatively acceptable, despite the fact that it is ungrammatical. If many transformations are needed

to reconstruct the VP, then the listener will have the impression that the sentence is very unacceptable.

To test this, Arregui et al. performed an acceptability judgment task examining VP ellipses with four types of antecedents, illustrated in (31).

- (31)
- a. None of the astronomers saw the comet, but John did.
 - b. Seeing the comet was nearly impossible, but John did.
 - c. The comet was nearly impossible to see, but John did.
 - d. The comet was nearly unseeable, but John did.

(31a) features a viable VP antecedent, while (31b) provides a nominalized gerund form of the verb phrase. In (31c), “the subject of the first clause is coindexed with an empty operator which binds an object trace,” while in (31d), the critical verb that should be used to reconstruct an antecedent VP is embedded in an adjective. By hypothesis, these conditions are thought to be presented in order of increasing difficulty of reconstruction, so the VP recycling hypothesis predicts that they are in decreasing order of acceptability. This prediction was born out in the acceptability judgment study.

While this system makes sense from the perspective of a listener confronted with a mismatch, it is worth asking why speakers would generate such sentences in the first place if they are indeed ungrammatical. One possibility offered by Frazier 2015 is that they are simply speech errors. It is tempting to propose is that speakers simply forget how the sentence began by the time they produce the second clause; it is already established, for example, that linguistic structures can be misremembered as having a different voice specification than is actually the case, with an added asymmetry that passives are more likely to be misremembered as actives than vice versa (Mehler, 1963; Arregui et al., 2006). On the other hand, it is less clear how this proposal accounts for mismatch sentences found in written sources that have been carefully edited.

Building on this Frazier and Clifton Jr. (2011) highlight another area where speakers are

systematically confronted with multiple possible structures that express the same meaning, namely the choice between conjunction of two verb phrases or conjunction of two clauses with co-referring subjects. They found that listeners readily “repaired” VP ellipses that had a syntactic blend of the two strategies as an antecedent, yielding roughly the same interpretation facts as when the ellipsis site had licit VP conjunction as the antecedent. They suggest that repair is particularly easy in cases that involve a syntactic blend that results from highly common encoding decision for speakers.

In a similar vein, Grant et al. (2012) note a possible role for QUD considerations in guiding repair operations. They note that many acceptable mismatch sentences, like (32), open with a clause that does not assert a particular state of affairs. They offer that such clauses may implicate that the state of affairs is not the case, which they call a *non-actuality implicature*, implicitly raising a QUD asking whether it is the case or not.

- (32) This problem was to have been looked into, but obviously nobody did.

They propose that this implicit QUD can guide interpretation of an elliptical clause, especially in cases where the antecedent needs to be repaired in order for the elliptical construction to be grammatical. In other words, the QUD explicitly raising the alternatives that the problem was or was not looked into facilitates the reconstruction of an active antecedent that is compatible with the ellipsis site.

Another approach to acceptable antecedent-ellipsis mismatches under a syntactic identity account involves the accommodation of alternative antecedents. Fox (1999, 2000) initiated this line with a refinement of Rooth’s (1992a) account of data like (33) and (34).

- (33) First John told Mary I was badmouthing her, and then SUE *heard I was badmouthing her*.

- (34) First John told Mary I was badmouthing her, and then SUE *did*.

In (33), the VP *heard I was badmouthing her* can be deaccented on the basis of the presence of *tell x I was badmouthing her* in the first clause. By contrast, the ellipsis site in

(34) cannot be interpreted as *hear I was badmouthing her*. Rooth argues that deaccenting (*phonological reduction*) and ellipsis are governed by a shared, semantically mediated constraint qualifying material for deaccenting when it is entailed by certain antecedent material. Ellipsis, but not deaccenting, is subject to an additional, syntactic constraint enforcing identity of form with the antecedent. Since *heard I was badmouthing her* is not syntactically isomorphic to the antecedent *tell x I was badmouthing her*, ellipsis (with the intended reading) is not licit in (34).

Fox's account aims to eliminate the differential licensing requirements for deaccenting and ellipsis. He proposes that both deaccenting and (VP) ellipsis are subject to parallelism requirements roughly comparable to the syntactic identity accounts discussed in broad terms above. What distinguishes the deaccenting in (33) from the illicit ellipsis in (34) is the presence of *accommodation-seeking material*.

Fox defines accommodation-seeking material as overt, non-F-marked material that is present in the reduced clause but not in the antecedent. The absence of F-marking on overt material (roughly analogous to a lack of pitch accent in English) marks the material as discourse-old or otherwise accessible. In cases where material is not F-marked, but is not accessible via syntactic identity with an overt antecedent, as the grammar requires by hypothesis, Fox argues that this serves as a *trigger* for accommodation of an alternative antecedent.

In the case of (33), the deaccented VP *heard I was badmouthing her*, or more minimally, the verb *heard*, is accommodation-seeking material because it is non-F-marked and not present in the linguistic antecedent. In essence, grammatical use of this material without F-marking requires the presence of the VP *heard I was badmouthing her* in the antecedent. Since *John told Mary I was badmouthing her* entails *Mary heard I was badmouthing her*, it is reasonable for a listener to accommodate this alternative linguistic structure, then proceed with interpretation as though this were the antecedent for the deaccented verb phrase.

By contrast, (34) does not contain any accommodation-seeking material. The only non-F-

marked material following the second-clause subject, *did*, can be construed as present in the antecedent (finite past tense, van Craenenbroeck 2012). As such, there is no indication to the listener that interpretation should proceed using anything but the original overt antecedent. The only available interpretation for the ellipsis site is *told Mary I was badmouthing her*, and an alternation to a structure using the verb *heard* is not supported. Fox notes that as a general rule, accommodation-seeking material in elliptical clauses should be extremely rare by virtue of the fact that it must be overt, whereas ellipsis by its nature entails missing material in the critical construction.

van Craenenbroeck 2012 aims to refine Fox’s proposal, in particular through examination of certain (sluicing) data points that appear to contradict Merchant’s 2001 Preposition Stranding Generalization shown in (35).

- (35) A language *L* will allow preposition stranding under sluicing iff *L* allows preposition stranding under regular wh-movement.

The offending data, from Spanish, are shown in (36) and (37). (36) shows that Spanish in general does not allow preposition stranding. In (37), however, the preposition *con* appears to have been stranded upon movement of *cuál* and elided.

- (36) *Qué chica rubia ha hablado Juan con?
 what girl blonde has talked Juan with
 ‘What blonde girl did Juan talk to?’
- (37) Juan ha habaldo con una chica rubia, pero no sé cuál.
 Juan has talked with a girl blonde but not I.know which
 ‘Juan has talked to a blonde girl, but I don’t know which.’

However, Rodrigues et al. (2009) and Vicente (2008) argue that the ellipsis site can be analyzed as a short cleft rather than a full wh-question, which would violate the preposition stranding generalization. This analysis is shown in (38).

- (38) Juan ha habaldo con una chica rubia, pero no sé cuál es ~~pro~~.
 Juan has talked with a girl blonde but not I.know which is it
 ‘Juan has talked to a blonde girl, but I don’t know which.’

The facts in Spanish conflict with those of Greek, which likewise does not allow preposition stranding, as shown in (39). As predicted by the preposition stranding generalization, (40) is ruled out; the accusative case on *pjon* indicates a stranded preposition inside the ellipsis site. However, (41) is also ungrammatical, despite nominative case on *pjos* that would be expected if the ellipsis clause were a short cleft.

- (39) *Pjon milise me?
 who.ACC she.spoke with
 ‘Who did she speak with?’
- (40) *I Anna milise me kapjon, alla dhen ksero pjon.
 the Anna spoke with someone, but not I.know who.ACC
 ‘Anna spoke with someone, but I don’t know who.’
- (41) *I Anna milise me kapjon, alla dhen ksero pjós.
 the Anna spoke with someone, but not I.know who.NOM
 ‘Anna spoke with someone, but I don’t know who.’

Based on these facts, along with data from Hungarian NP ellipsis, Hebrew V-stranding VP ellipsis, and subject extraction in Hungarian, van Craenenbroeck proposes a modified characterization of antecedent accommodation consisting of two assumptions, shown in (42) and (43).

- (42) An accommodated antecedent can only be built up from non-F-marked overt material present in the discourse or from elements that are freely available in any discourse.
- (43) Material that is extracted from an ellipsis site cannot be part of an accommodated antecedent.

The first assumption (42) enables the accommodation of a short cleft structure in (38). This is because van Craenenbroeck assumes that pronouns and copulas are freely accommodable in any discourse, following Merchant’s (2004) observation that these can generally be elided even in discourse initial contexts (e.g., *Some guy she met at the park* in response to a friend’s puzzled look after spotting a new person at a party).

The second assumption (43) blocks this technique in cases like (41). The critical generalization is that the copula strategy is blocked in languages that mark case, unless the case assigned by the preposition and the case in a copular clause are accidentally identical (Polish instrumental), the two cases are syncretic, or case marking can be dropped. In other words, the copula strategy is blocked any time overt case marking appears in the ellipsis clause that does not match the case assigned by the preposition. By blocking non-pronoun, non-copula material that has been extracted from the ellipsis site from forming an accommodated antecedent, van Craenenbroeck rules out the possibility of the copula analysis for the bulk of case-marking languages.

Finally, Thoms (2015) proposes that a syntactic identity requirement for ellipsis can be augmented with the generation of algorithmically-defined alternative antecedents in order to generate the correct distribution of acceptable versus unacceptable antecedent-ellipsis mismatches. Thoms notes that mismatches between antecedent and ellipsis site are not uniformly prohibited. He suggests that the key difference distinguishing acceptable and unacceptable mismatches is that in unacceptable cases, a variable acts as an antecedent for a non-variable (in particular, he considers cases where a trace is the antecedent for a non-trace). As an example, he points to the Scottish Gaelic data in (44) and (45):

- (44) A bheil thu a' faighinn leabharaichean bho Fheargais an-dràsta?
 C-Q be-PRES-IND 2SG IMP get-VN books from Fergus just-now
 ‘Are you getting books from Fergus just now?’

Chan eil $\emptyset_1$, ach gheibh $\emptyset_2$ a-màireach!
 C-NEG be-PRES-DEP but get-FUT-IND tomorrow

‘I’m not, but I will tomorrow!’

[$\emptyset_1$ =‘getting books from him’]

[$\emptyset_2$ =‘get books from him’]

- (45) An dh’fhuair thu leabharaichean bho Fheargais?
 C-Q get-PST-DEL 2SG books from Fergus
 ‘Did you get books from Fergus?’

*Cha dh'fhuair $\emptyset_1$, ach tha $\emptyset_2$ an-dàsta!
 C-NEG get-PST-DEP but be-PRES-IND now

‘I didn’t, but I am now!’

[$\emptyset_1$ =‘get books from him’]

[$\emptyset_2$ =‘getting books from him’]

Thoms’ general syntactic analysis of Scottish Gaelic argues that initial lexical verbs are extracted from lower in the structure. Thus, in (44), the elliptical clause containing $\emptyset_2$ has an extracted lexical verb. (45) shows the opposite pattern, with an extracted lexical verb in the antecedent. (44) is claimed to be acceptable because an extraction from an ellipsis site does not require a parallel extraction from an antecedent. (45) is ungrammatical because an extraction from an antecedent requires a parallel extraction from the ellipsis site, which does not hold in (45).

Having established what he claims is the locus of unacceptable mismatches, Thoms then seeks to articulate a mechanism by which acceptable mismatches are derived. He proposes that the ellipsis identity relation is one of strict syntactic identity, but is “loose” in the sense that the relation can exist between an ellipsis and an accommodated, rather than overt, antecedent. Accommodation is triggered by any mismatch between the ellipsis and the overt antecedent, and is modeled as proceeding by the construction of a set of structurally-defined alternative antecedents (following Katzir 2007). The set of possible alternative antecedents consists of those structures that can be produced through a finite number of applications of the operations of *deletion*, *contraction*, and *substitution*, and that are semantically isomorphic to the overt antecedent. By definition, the alternative structures are “at most as complex” as the original overt antecedent. The trace-mismatch data are accounted for by the requirement that a simple node, i.e., variable, cannot be replaced by a complex node specified for lexical content, according to the accommodation algorithm.

Fox’s, van Craenenbroeck’s, and Thoms’ analyses each provide for the accommodation of an alternative antecedent when an elliptical clause violates the hypothesized grammatical

requirement for syntactic identity with an overt antecedent. Importantly, these mechanisms were proposed to account for relatively minor formal mismatches between ellipsis clauses and their overt linguistic antecedents, and the types of accommodation that are possible are hypothesized to be minor and highly constrained by the linguistic material that is available in the sentence. (It is also worth noting again that much of the critical data is from sluicing rather than VPE).

Nevertheless, it appears that there is independent theoretical motivation for the accommodation of alternative antecedents in VPE sentences, and it is worth asking whether an accommodation or repair mechanism might play a role in integrating information from the discourse context – the nonlinguistic context as well as the linguistic discourse outside a local antecedent – in the interpretation of VPE. As will be discussed in more detail below, this type of account provides a possible alternative to semantic identity models of VPE in cases where the broader context is brought to bear, such as antecedentless/exophoric ellipsis. It may be possible to maintain a syntactic-identity grammar of VPE, while a supplemental accommodation or repair mechanism accounts for some deviation from the grammatically supported interpretation in the context of a linguistic antecedent, as well as for the possibility of interpretation in the absence of a linguistic antecedent.

2.5 Antecedentless VP ellipsis

The possibility of VP ellipsis sites receiving an interpretation in the absence of a linguistic antecedent – called *exophoric* or *antecedentless* VP ellipsis – was mentioned above as possible evidence for semantic identity between ellipsis sites and some meaning available in the discourse. Indeed, the interpretation of such instances of VPE is straightforwardly handled by a semantic account, provided that some antecedent meaning is sufficiently salient in the discourse at large, whereas it is less apparent how a syntactic identity model can handle cases with no useable prior linguistic structure. (It is also worth noting that while interpretation is handled relatively easily by a semantic identity account, it not obvious what factors license

antecedentless VPE, as it is quite rare.)

Antecedentless VPE has received relatively little attention in the literature on ellipsis, although it has received some additional treatment in recent years. This dissertation considers whether additional consideration of antecedentless ellipsis – as well as ellipsis in contexts inspired by known antecedentless examples – can inform the identity question for ellipsis, and perhaps lead to a unified account of ellipsis interpretation across antecedentless and canonical cases of VPE. As such, it is worth considering the prior literature on antecedentless VPE in some detail.

The question of whether verb phrase ellipsis structures can receive (part of) their interpretation from the nonlinguistic context dates at least to Hankamer and Sag (1976). Hankamer and Sag propose that all anaphoric processes are either *deep anaphors* or *surface anaphors*. Deep anaphors, like pronouns, are inserted into a syntactic derivation at a very early stage and are not fundamentally affected by later transformations. Surface anaphors, including verb phrase ellipsis, are fully syntactically articulated early in the derivation, but on the surface their structure is rendered opaque by intervening transformations such as deletion.

An important distinction between deep and surface anaphora proposed by Hankamer and Sag is the mechanism by which they receive their interpretation. Deep anaphora can be used endophorically (i.e., anaphorically or cataphorically with respect to a syntactically present antecedent - “syntactic control”) or exophorically (i.e., in the absence of an overt antecedent, possibly receiving an interpretation based on inference from the nonlinguistic context - “pragmatic control”). By contrast, surface anaphors are suggested to require a syntactically identical antecedent in order to receive their interpretation; that is, surface anaphors like verb phrase ellipsis require syntactic control, and cannot receive any part of their interpretation from the nonlinguistic context. (Note that Sag and Hankamer (1984) later revised their terminology to reflect the contribution of a discourse model, rather than the nonlinguistic context itself, to the interpretation of deep anaphors. They propose renaming deep anaphora *model-interpretive* anaphora and surface anaphora simply *ellipsis*.)

In support of this claim, Hankamer and Sag present several examples, such as (46), that appear to demonstrate that elided verb phrases cannot receive an interpretation in the absence of a syntactic antecedent.

- (46) [Context: Sag raises a cleaver and prepares to hack off his left hand.]
- a. *Hankamer*: Don't worry, he never actually does it.
 - b. *Hankamer*: #Don't worry, he never actually does.

In (46a), the deep anaphor *do it* freely derives its interpretation from the unspoken context, in which it is clear that Sag is about to hack off his hand. In (46b), though, the elided verb phrase *hack off his left hand* is apparently not available despite the salience of such a concept in the discourse context.

Schachter (1977) responded to Hankamer and Sag with a collection of naturally occurring counterexamples of verb phrase ellipsis for which the interpretation came from the nonlinguistic context, including (47) and (48).

- (47) [The band starts playing and several couples head for the dance floor. Pulling back his chair and half-rising, John says to Mary:]
- Shall we?

- (48) [In a hair color commercial:]
- Does she or doesn't she? Only her hairdresser knows for sure.

On the basis of these examples, Schachter argues that the empirical generalization proposed by Hankamer and Sag that VPE cannot receive its interpretation from information available in the nonlinguistic context. Schachter (1978) further argues for a “propredicate” analysis of VPE that unifies its interpretation with that of other anaphoric expressions such as pronouns, allowing VPE in principle to receive its interpretation from either the linguistic or nonlinguistic context according to the same principles as pronouns.

Hankamer (1978) responds by saying that Schachter's examples are the exceptions that prove the rule. He suggests that acceptable instances of exophoric verb phrase ellipsis are

possible only in exceptional, independently motivated circumstances. He suggests that constructions like (47) are idioms, and cannot take on other meanings besides the ones they have become conventionally associated with. Similarly, (48), where the elided verb phrase is interpreted as *color her hair*, is parasitic on the version of *Does she or doesn't she?* where the elided VP is interpreted as *screw*, itself probably an idiom generated by the principle of taboo avoidance. Insofar as Schachter failed to produce an example of exophoric VP ellipsis that could not be explained in such a manner, Hankamer concluded that the generalization that ellipsis cannot be pragmatically controlled was correct.

Pullum (2000) supports the intuition reported by Hankamer that the small number of exophoric cases of verb phrase ellipsis compared to the massive number of possible elliptical structures indicates that acceptable cases are idiomatic. Pullum calculates that there are over two thousand possible verb phrase ellipsis constructions with pronominal subjects, accounting for possible combinations of auxiliaries, polarity, and subjects. Nevertheless, Pullum claims to have found only about fifteen combinations that have acceptable exophoric uses, and these only in particular contexts. He agrees with Hankamer that in light of the extremely low proportion of possible sentences that are actually acceptable, the best generalization is a blanket ban on exophoric ellipsis with exceptions for those few structures that actually do receive their interpretations from the nonlinguistic context.

However, subsequent investigation has revealed that antecedentless VPE is more widespread, and more productive, than concluded by Hankamer and Pullum. For instance, Merchant (2004) provides several examples, such as (49), in which an apparent VPE site receives a non-idiomatic interpretation from a source other than an overt VP antecedent.

- (49) [Harry, alone in a corridor, discovers a classmate in an enchanted paralysis on the floor. Just then, the evil groundskeeper chances upon him, and, assuming Harry has laid the spell, runs to fetch a teacher. In a moment, he returns with the teacher, who shakes her head and turns away. Harry, aghast at being suspected of the evil deed, calls after her:]

I swear I didn't!

(*Harry Potter and the Chamber of Secrets*, 2002)

Despite the acceptability of such examples, however, Merchant does not conclude that VPE as such can receive an interpretation directly from a nonlinguistic context, at least in the sense of the nonlinguistic context providing a specific syntactic structure with specific lexical items to be inferred as present in the ellipsis site. Rather, he suggests that broader discourse contexts are capable only of making the verbal anaphor *do it* manifest. With a sufficiently supportive discourse context, interpreters who encounter an apparently antecedentless instance of VPE infer that the elided content is *do it*, and interpretation proceeds according to the typical mechanism for *do it*. Importantly, *do it* is known to be flexible in its ability to receive an interpretation not only from a local linguistic antecedent, but also from information generally available in the discourse context at large.

Miller and Pullum (2013) likewise note an expanded distribution of productive exophoric VP ellipsis, noting several examples taken from the Corpus of Contemporary American English, including (50).

- (50) The aisles at the Lakewood Wal-Mart are surprisingly packed at 11 p.m. “Can we? Can we?” Vanessa tugs at her mother, pointing to a rack of “Lady and the Tramp” DVDs. Diaz shrugs. OK.

Miller and Pullum propose a unified account of exophoric and canonical endophoric VPE expressed in terms of the discourse requirements imposed on the use of VPE. Importantly, they argue that VPE does not require identity with a linguistic antecedent at all, but is actually freely available in any discourse that meets these requirements.

Building on Kertz’s (2013) approach, Miller and Pullum distinguish two different types of VPE construction. In the first, *auxiliary-choice* ellipsis, the subject of the VPE clause is identical to the subject of the antecedent, while the auxiliary is focused. They propose that this type of ellipsis is licensed any time “a choice between the members of a jointly exhaustive

set of alternative situations [is] highly salient in the discourse context, and the point of the utterance containing the [VPE] is strictly limited to selecting one member of that set.” In other words, if the context, linguistic or otherwise, makes salient a set of propositions that contrast in tense, aspect, modality, or most frequently, polarity, and the point of the VPE utterance is to select one of the options as being the case, VPE is licensed.

The second type of VPE is called subject-choice ellipsis. In this case, the subject of the elliptical clause is distinct from that of the antecedent. This type of ellipsis is licensed when a property is highly salient in the discourse, and the elliptical utterance serves to identify someone or something that possesses the property.

A defining feature of Miller and Pullum’s approach is that it does not specifically require a linguistic antecedent for VPE. Rather, the “antecedent,” following Cornish (1999), is a representation of the discourse that enables interpretation of the ellipsis site. The linguistic structures treated as “antecedents” in previous literature are merely “antecedent-triggers,” or linguistic objects that directly contribute to the construction of the antecedent (Miller and Pullum, 2013; Miller and Hemforth, 2014).

Linguistic antecedent-triggers can contribute to the construction of an antecedent, but they are not necessary for the successful formation of an antecedent. Miller and Pullum argue that the apparent need for either syntactic or semantic identity with a linguistic antecedent reported in the prior literature is actually an illusion; linguistic identity is actually not required, but it appears to be necessary because only relatively few nonlinguistic contexts can satisfy the discourse requirements on the felicitous use of VPE.

Focusing on auxiliary-choice VPE, Miller and Pullum note four ways prior linguistic material can make a set of alternative situations sufficiently salient to license ellipsis: assertions, polar questions, requests for permission, and directives (imperatives, or comments on the advisability or permissibility of some course of action) all count. They argue that nonlinguistic contexts are altogether incapable of making assertions or asking polar questions. By contrast, they claim that the nonlinguistic context can supply a salient alternative set in

cases where the VPE utterance constitutes a request for permission or a directive. (50) is an example of the former, while (51), also from COCA, is a case of the latter.

- (51) Mabel shoved a plate into Tate’s hands before heading for the sisters’ favorite table in the shop. “You shouldn’t have.” She meant it. The sisters had to pool their limited resources just to get by.

Miller and Pullum claim that when the discourse conditions on the use of VPE are met, ellipsis is freely available and readily interpretable. Along with Miller and Hemforth (2014), who propose a similar account to deal with VPE utterances with nominal antecedents, they propose an identity-of-meaning account where VPE is interpreted with respect to a holistic representation of discourse that takes into account both linguistic and nonlinguistic information rather than with respect to a linguistic antecedent. They propose that the apparent need for a linguistic antecedent emerges because nonlinguistic contexts can only rarely satisfy the discourse requirements for VPE use. Likewise, they suggest that an apparent preference for syntactic identity stems from the heuristics interpreters use to identify antecedent triggers, as the processor can access an antecedent supported by a recent syntactic antecedent-trigger more readily than one constructed through consideration of the discourse at large.

2.6 An empirical gap

The observation that verb phrase ellipsis can be judged acceptable and used productively even in the absence of a linguistic antecedent verb phrase adds yet another dimension to the ellipsis identity question. On its surface, this finding appears to support semantic or discourse-based accounts of the ellipsis identity relation, as it is unclear how syntactic identity accounts would handle the fact that VPE can be interpreted in the absence of any prior linguistic structure. Indeed, several researchers have argued that exophoric VPE is evidence in favor of such accounts (Schachter, 1978; Miller and Pullum, 2013; Kehler, 2019).

On the other hand, it may be possible to reconcile the availability of exophoric VP

ellipsis with syntactic accounts of ellipsis interpretation. The most plausible such account is one like Merchant’s (2004), which effectively assigns a special interpretation mechanism (inferring that the elided phrase is *do it*) only to exophoric ellipses. Another possibility would be to suggest that ellipsis requires a syntactically isomorphic antecedent, and that exophoric VP sites trigger construction of a covert antecedent, in the vein of the accommodation and repair accounts discussed above (Fox, 2000; Arregui et al., 2006; van Craenenbroeck, 2012; Thoms, 2015). Allowing such mechanisms would raise the question of why the distribution of exophoric VPE is so limited, but they are not outside the realm of possibility.

Experimental investigation of exophoric VPE may lead to valuable insights as to its mechanism of interpretation. This dissertation undertakes such investigation, but also considers another type of evidence that is directly inspired by previous work on exophoric ellipsis.

Work on canonical VP ellipsis has largely focused exclusively on exclusively linguistic discourses – for example, active ellipsis clauses with single passive antecedent VPs. This is not a shortcoming of this research, but a natural consequence of the fact that even the relationship between ellipsis sites and prior linguistic material is quite complex. Even examples which consider a “broader” discourse than a local antecedent VP, such as Webber’s (1978) split antecedent cases or Hardt’s (1990) corpus examples with multiple clauses intervening between ellipsis sites and their antecedents, generally draw from exclusively linguistic discourses.

Likewise, the literature on antecedentless ellipsis largely features examples where the ellipsis site is interpretable using information from an exclusively nonlinguistic context, where there is no linguistic antecedent VP available at all. Again, this is not a flaw with the research, since up until this point much of the focus has been to determine whether antecedentless ellipsis was even possible.

These observations point to a possible data source that has previously been overlooked in the literature on the ellipsis identity question: complex discourse contexts in which there is both a linguistic VP antecedent as well as salient information in the broader (e.g., nonlin-

guistic) discourse context that could potentially contribute to interpretation. Such contexts should shed light on the question of whether interpretation of the ellipsis site, canonically speaking, is exclusively syntactic, or whether semantic and discourse considerations play a role as well. Unlike completely exophoric contexts, these complex contexts contain a viable linguistic VP antecedent that could be accessed by a syntactic interpretation mechanism. Thus, if the interpretation of ellipsis sites in these contexts does not proceed completely according to the linguistic antecedent, it suggests that semantic or discourse factors play a role as well. A possible example of such a context is shown in (52).

- (52) [A father and child are in a grocery store. The child places five candy bars in the shopping cart. This is done in an exaggerated and ostentatious manner, making it clear that the child wants to buy precisely five candy bars.]

Child: I want to buy candy bars!

Father: We can't.

The fundamental question in this context is whether the ellipsis site is interpreted as making any comment about the specific number of candy bars that can be purchased. On the one hand, the nonlinguistic context supports the notion that the number of candy bars is at-issue, with the child specifically interested in purchasing five candy bars. If the father's elliptical response can turn on this information, then it might (among other interpretations) mean something like 'We can't buy five candy bars [but maybe we could buy fewer].' On the other hand, the overt antecedent VP, *buy candy bars*, does not contain a numeral. A syntactic identity account of VPE interpretation predicts that the only interpretation that should be available is the one that arises from considering the sentence with the antecedent VP *in situ*, *We can't buy candy bars*. The only reasonable interpretation is that no candy bars can be purchased. Thus, if listeners in this context find that the interpretation where the number of candy bars is at-issue is available under ellipsis, it supports some role for semantic or discourse information in the ellipsis interpretation mechanism.

The broader literature on the ellipsis identity relation gives rise to several different pos-

sible mechanisms of interpretation in complex discourse contexts, summarized below.

Syntactic identity with “escape hatch” for exophoric cases: One possibility is that canonical ellipsis is interpreted under strict syntactic identity with a linguistic antecedent when one is available, and that exophoric ellipsis is interpreted using a completely different mechanism. In other words, exophoric ellipsis is “defective” and exempt from the normal interpretation process. This separate mechanism may be inference of an elided VP anaphor like *do it*, per Merchant (2004), or some other mechanism; at any rate, it is clear that interpretation in exophoric contexts must proceed entirely according to the information available in the discourse at large, taking into account nonlinguistic information as well as, for example, linguistic material that is not viable as a VP antecedent.

Syntactic identity with accommodation or repair: A second possibility is that the grammar of VPE mandates a syntactically isomorphic antecedent, but that alternative antecedents can be constructed under certain conditions. For exophoric VPE, this would suggest that information from the nonlinguistic context supports accommodation of a covert linguistic VP, which is then used to interpret the ellipsis site. However, Merchant (2004) raises important concerns about the overly wide range of VPs that contexts might make available versus the relatively restricted interpretations available for exophoric VPE.

In complex discourse contexts, the predictions of this account are somewhat less straightforward. Prior accommodation and repair accounts rely on a formal mismatch between the ellipsis and antecedent clauses to trigger the operation. Such a mismatch is clear in exophoric cases, but not in complex discourse contexts that feature a viable antecedent VP. Thus, in its simplest form, this account predicts that interpretation in such contexts should proceed entirely according to the content of the antecedent VP. However, the next chapter will discuss the possibility of alternative, discourse-driven triggers for such an operation, which opens the door for the consideration of alternative meanings while maintaining a syntactic identity-based grammar.

Discourse-driven interpretation with simple heuristics: Semantic identity and

discourse-driven accounts of VPE interpretation often leave the exact mechanisms by which an antecedent meaning is selected during interpretation unspecified. For instance, Miller and Hemforth (2014) appeal to processing heuristics to explain the apparent preference for interpretation under identity with a linguistic antecedent, but they do not explore this in much more detail than to say that recent syntactic objects may be more cognitively accessible than more articulated representations of the discourse at large.

One possible extension to this approach is that the interpretation heuristics are actually quite simple – VPE is interpreted simply by identifying whatever compatible meaning is most salient or cognitively available in the discourse. This would allow for the interpretation of exophoric VPE, since ellipsis sites can freely access information from the nonlinguistic discourse. In complex discourse contexts, this account predicts a straightforward mapping between measures of the availability or salience of various propositions in the context prior to the use of a VPE utterance and their feasibility as interpretations for a VPE site. To the extent that the availability of each interpretation is influenced by information from the nonlinguistic or broader linguistic context, then this information is also available in ellipsis interpretation. Listeners might categorically select the most available meaning, or proportionally consider multiple meanings according to their relative availability.

Discourse-driven interpretation with complex heuristics: A second possibility for the discourse-driven approach is that the interpretation heuristics go beyond simply accessing the most salient interpretation. Hardt (1993) outlines some heuristics that he proposes govern the selection of an antecedent meaning from an articulated linguistic discourse. These include a preference for subject co-reference between the antecedent and ellipsis clauses, a preference for antecedents located in clauses related to comparative or adverbial clauses, and a general preference for the antecedent to be local to the ellipsis. Similarly, work on anaphoric constructions like pronouns has revealed that the expression itself can introduce complex, form-specific biases as to what information should be considered during interpretation (Gundel et al., 1993; Kehler and Rohde, 2013).

Thus, a discourse-driven account with more articulated processing heuristics might predict what appears to be a “mix” of syntax- and discourse-driven interpretation of VPE sites situated in complex discourse contexts. In particular, a heuristic that favors, say, local linguistic antecedents might lead to viable antecedents exerting an outsize influence on the distribution of available interpretations that goes beyond their contributions to discourse salience in the context at large.

The major focus of the next chapter is experimentally assessing the interpretation of VPE in complex discourse contexts like the one shown in (52). The experimental results effectively rule out a strictly syntax-driven account of VPE interpretation, but are likewise incompatible with the “simple” discourse hypothesis. Thus, the investigation constrains the possible interpretive mechanism from both the purely syntax and purely meaning sides of the spectrum. The discussion focuses on possible paths forward for the syntax-and-accommodation approach and the “complex” discourse approach.

CHAPTER 3

VERB PHRASE ELLIPSIS INTERPRETATION IN COMPLEX DISCOURSE CONTEXTS

The previous chapter introduced several possible mechanisms for context-situated VPE interpretation: a syntactic identity mechanism with a separate “escape hatch” for interpreting antecedentless VPE, a syntactic identity mechanism with somewhat free construction of covert alternative antecedents, a simple discourse-driven model that considers interpretations in proportion with their prior salience in the discourse, and a more complex discourse-driven model with several interacting heuristics for identifying the interpretation from the prior context. These accounts sit on a spectrum from strongly syntactically driven (strict identity) to strongly driven by discourse salience (simple discourse model).

“Complex discourse contexts,” where there is both a viable linguistic antecedent and salient information in the broader discourse context, were highlighted as a possibly informative empirical testing ground for these hypotheses. The hypotheses make distinct predictions for interpretation in these contexts. A strict syntactic account predicts that interpretation should proceed according to the form of the linguistic antecedent, when present and viable, and that information from the broader context should be ignored. The simple discourse account predicts that the distribution of interpretations available for an ellipsis site should mirror the availability of the corresponding propositions in the prior discourse. The less restricted identity account and the complex discourse account leave open the possibility of more complex interpretation patterns; in particular, it is possible that interpretation would be similar to the predictions of the identity account, but with a slight shift toward the predictions of the simple discourse account, or vice versa.

The present chapter opens the investigation of VPE interpretation in complex discourse contexts. The experiments presented below feature a novel paradigm that elicits interpretations of verbal anaphors in complex discourse contexts. In each experiment, both the

linguistic antecedent and the broader discourse context vary in the extent to which they make number information about the patient of a particular action at-issue or not at-issue. Participants responded to various prompts asking about the interpretation of target sentences, including VPE sites or other constructions of interest, in these discourse contexts to probe whether they interpret number information as included in the meaning of the target. Thus, the paradigm can be used to determine whether only information from the linguistic antecedent or information from the broad (linguistic and nonlinguistic) discourse context affects the interpretation of the target, or whether the interpretive mechanism hybridizes information from these sources.

The strict identity model and the simple discourse-driven model make the clearest predictions for interpretation in complex discourse contexts, so in the body of this chapter, the experiments are framed in terms of their consequences for these two accounts. For convenience, these are referred to as the “identity” and “discourse” models, respectively. Unsurprisingly, the facts of interpretation in context are more complicated than is predicted by either of these two models, and at the end of the chapter, the less restricted identity model and the complex discourse model are proposed to be the remaining viable accounts.

Experiment 1 elicits ratings for a number-unmodified and a number-modified interpretation for verb phrase ellipsis in the variable discourse context described above. The results show that even in the presence of a viable linguistic antecedent, the distribution of interpretations available for a VPE utterance is sensitive to the information available in the broader discourse context. This is problematic for a strict formulation of the identity account of VPE interpretation, as variations in the broader discourse context should not affect interpretation when the antecedent is held constant according to this model.

Experiment 2 tests whether the contextual effect found in Experiment 1 is unique to ellipsis, or is merely a broader effect of utterance interpretation in a highly informative discourse context. It does this by replacing the critical VPE utterance from Experiment 1 with a sentence with a fully realized VP that is identical to the VP from the antecedent

utterance. The results show that the context effect from Experiment 1 is largely unique to ellipsis, further indicating that the strict identity account cannot account for the data from that experiment.

Experiment 3 assesses the discourse coherence of the number-unmodified and number-unmodified interpretations from Experiment 1 in each discourse context. The goal of this experiment is to determine whether a simple discourse model can account for the data from Experiment 1. The results show that there is not a straightforward mapping between the meanings that are most coherent in each discourse context and the interpretations available under ellipsis, suggesting that it is unlikely a simple discourse model can adequately account for VPE interpretation in complex discourse contexts.

Finally, Experiment 4 extends the experimental paradigm to another verbal anaphor, *do that*. Unlike VPE, which is identified as a “surface” anaphor in the theoretical literature, *do that* is identified as a “deep” anaphor, meaning it can freely receive its interpretation from both a linguistic antecedent or the broader discourse context (Hankamer and Sag, 1976). By hypothesis, then, *do that* should be much more sensitive to experimental manipulations in the broader context than VPE. Experiment 4 thus serves both as a check on the validity of the experimental paradigm and as an initial extension of the paradigm to context-dependent verbal expressions beyond VPE. The results show that *do that* is indeed more sensitive to manipulations in the broader discourse context, but in light of the results for VPE, there is a possibility that the two expressions lie on a cline of sensitivity to the broader discourse rather than on two sides of a categorical divide.

Considered together, the experimental results suggest that neither a strict identity account nor a simple discourse account can adequately capture the facts of VPE interpretation in complex discourse contexts. The distribution of interpretations available for an ellipsis site is driven by the information that is available in the broad discourse context, including both linguistic and nonlinguistic information, but at the same time is constrained by the content of the linguistic antecedent to an extent beyond its simple contribution to discourse salience.

This indicates that both linguistic identity and discourse information are implicated in a hybrid interpretation mechanism for VPE, such as a less restricted identity account or a complex discourse account. These models are revisited in the general discussion and receive further treatment in the next chapter.

3.1 Experiment 1: VPE interpretation in context

The goal of this experiment is a first assessment of the interpretation of verb phrase ellipsis in complex discourse contexts using the novel experimental paradigm.¹ The identity and discourse accounts of VPE interpretation make different predictions for how interpretation proceeds when there is both a viable linguistic antecedent and salient information in the nonlinguistic context. However, the data made available in the literature so far makes it difficult to determine whether the predictions of one or the other model are actually borne out in such contexts. The present experiment addresses this empirical gap by eliciting ratings of number-unmodified and number-modified VPE interpretations in scenarios where both the linguistic antecedent and the broader discourse context vary in the extent to which they make number information at-issue.

3.1.1 *Design and materials*

The main task in the experiment was for participants to view a discourse consisting of a comic strip representation of a context, an optional antecedent utterance made by one of the characters visible in the comic strip, and a VPE reply utterance made by the other character. Following this discourse, participants were prompted with one possible interpretation of the VPE reply utterance and had to rate its likelihood as the character’s intended meaning on a seven-point Likert scale. A screenshot of a critical experimental item is shown in Figure 3.1.

The experiment featured a 3 (comic strip context) x 3 (antecedent) x 2 (interpretation

1. Experiments 1 and 2, with preliminary analysis and discussion, were previously published in conference proceedings form as Geiger and Xiang (2017).

In an office



Standing worker: I'm going to take four candies!

Sitting worker: You can't.

On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the sitting worker meant:

You can't take any candies.

(Unlikely)

(Likely)

Use number keys or click boxes to answer.

Figure 3.1: Sample screen from Experiment 1

prompt) design. Both the comic strip context and the linguistic antecedent varied primarily in the extent to which they made number information about an inanimate referent in the discourse context salient, as described below. The interpretation prompts varied between probing the likelihood of a number-unmodified or number-modified interpretation of the ellipsis site.

The three levels of comic strip context were *unavailable*, *available*, and *salient*. The intent behind these levels is that they form a continuum of availability for a critical piece of number information, with this information least salient in the unavailable strip and most salient in the salient strip. Sample comic strips for one experimental item are shown in Table 3.1. Each scenario was associated with a critical numeral that was potentially at issue. For instance, in the item shown in Table 3.1, the numeral at issue was *five*.

The *unavailable* strip features no visible interaction between the characters and the potentially modified referent. For instance, in the comic shown, neither the boy nor the father interact with any candy bars, and the critical numeral is “unavailable.” In the *available* strip, a character interacts with the critical number of referents, but not in an ostentatious way. In the strip shown, the boy takes five candy bars and places them in the cart. The fact that there are exactly five candy bars is not made highly salient, but could be verified upon reinspection, so the number information is “available.” Finally, in the *salient* strip, the character interacts with the critical number of referents in an ostentatious way, manipulating them one at a time and counting them out verbally. The intent of this level is that the numeral information be highly salient.

One level of the antecedent manipulation, *exophoric*, is actually the absence of an antecedent, with no utterance attributed to the relevant character (e.g., the son in Table 3.1). This condition served both as a check that participants were sensitive to the salience manipulation intended in the comic strip contexts, and an empirical investigation of the claim that exophoric VPE is straightforwardly resolved via discourse reference (e.g., Miller and Pullum, 2013). The other two conditions varied the availability of linguistic numeral information through the presence or absence of a numeral in an antecedent describing an intended course of action concerning the referent. The *unmodified* antecedent contained no numeral, and the *modified* antecedent was identical to the unmodified antecedent with the critical numeral added to the direct object as shown in Table 3.1.

The VPE reply utterance was a response attributed to the character in the comic strip who did not utter the antecedent. In all cases, this was an utterance with negative polarity denying permission for or the advisability of the course of action described in the antecedent, e.g., *We can't* or *You shouldn't*. These utterances were chosen because Miller and Pullum (2013) argue that VPE utterances constituting comments on permissibility or advisability are the most likely to be acceptable when used exophorically. Thus, by hypothesis, such utterances should sound relatively natural to participants in the exophoric condition, and

Unavailable comic



Available comic



Salient comic



Exophoric antecedent	[no utterance shown]
Unmodified antecedent	<i>Son:</i> I want to buy candy bars!
Modified antecedent	<i>Son:</i> I want to buy five candy bars!
VPE reply	<i>Father:</i> We can't.
Polar prompt	On a scale from 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the father meant: We can't buy any candy bars today.
Scalar prompt	On a scale from 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the father meant: We can't buy five candy bars today, but maybe we could buy fewer.

Table 3.1: Sample stimuli for Experiment 1

may also be more likely to show context effects in the conditions with a linguistic antecedent, as well.

Following the VPE reply, participants read a prompt eliciting a rating of the likelihood that the character uttering the VPE reply intended one of two interpretations. The first interpretation that could be elicited was one according to which the suggested course of action involving the referent could or should not be performed at all, regardless of number information. This is called the *polar* interpretation prompt since it deals with a binary choice between taking and not taking an action, e.g., *We can't buy any candy bars*. The second prompt elicited likelihood ratings for an interpretation that denied only the possibility or advisability of taking the action with a particular number of referents, but explicitly ruled in the possibility of taking the action with a lower number of referents. This is called the *scalar* interpretation prompt because it denies only a proposed position on the numeral scale, but not the action itself, e.g., *We can't buy five candy bars today, but maybe we could get fewer*.

The experiment included 6 critical items following the design outlined above. In addition, there were 10 filler items. The fillers involved a similar task of viewing a comic strip context, reading a discourse, and providing a likelihood rating for an interpretation of a VPE utterance, but did not feature the critical numeral information. The full set of critical and filler materials are shown in Appendix A.

The critical items were organized into lists so that each participant saw six different conditions for the six critical items, including seeing each level for the antecedent manipulation twice, each level for the comic strip context twice, and each level for the interpretation prompt three times. The script controlling the experiment was written so that the experimental list changed only when a participant completed the experiment; this resulted in a somewhat uneven distribution of observations across the experimental conditions and across the experimental items, but each condition was rated at least 35 times.

3.1.2 Procedure

Participants were recruited on the web-based Amazon Mechanical Turk (AMT) platform, where they were told they needed to be native English speakers aged 18 or over to participate in the experiment. Upon choosing to participate in the experiment on AMT, participants followed a link to the experiment, which was hosted on Ibex Farm (Drummond, 2016). Upon following this link, participants provided informed consent and completed a demographic questionnaire before the experiment began.

After the demographic questionnaire, participants viewed the instructions for the experiment. They were told to study the comic strip context carefully, including any details about the context, then silently read the accompanying antecedent and reply as well as the interpretation prompt. They were then instructed that they would view a Likert scale from 1-7 below the interpretation prompt, and that they could use this scale to answer the prompt either by clicking the numbers shown on the screen or by using the corresponding number key on their keyboard. Participants were told not to overthink their answers, and to follow their first instinct in cases where they felt unsure of their answer.

After viewing the instructions, participants completed two practice items. These items were designed to familiarize the participant with the layout of the experiment. They each featured a comic strip context and utterances plus an interpretation prompt asking the participant for the likelihood of a particular interpretation for one of the utterances. One practice item asked for the likelihood of a very likely interpretation, and the other for an unlikely interpretation. After each practice item, there was a feedback screen telling the participant that they “probably” selected a high (5-7) or low (1-3) rating for the likely and unlikely interpretation, respectively.

After the practice items, participants completed the 6 critical items and 10 fillers in a random order. Following this, they entered their AMT Worker ID and received a code to enter on the AMT survey to signal completion of the experiment. After successfully entering the completion code, participants received payment through the AMT platform.

3.1.3 *Participants*

195 participants participated in the experiment. 113 were female and 82 were male. The mean age of the participants was 35.7 years.

For every web-based experiment reported in this dissertation, participants were informed on AMT that being a native speaker of English was a prerequisite for participation in the experiment. Nevertheless, several participants failed to self-report as native speakers of English on the Ibex Farm demographic surveys. This includes participants who listed only languages besides English as their native language, participants who indicated that English was their native language in a language other than English, and participants who mistakenly answered a different demographic question (e.g., provided their age) in the box intended for native language information. For Experiment 1, 3 participants failed to identify themselves as native speakers of English and their responses were excluded from the analysis.

The design of the Ibex Farm experiment did not include any minimum response time or other checks that participants were paying attention. As such, some AMT users completed the experiment merely by pressing random number keys as quickly as possible to advance through the screens to the end of the experiment. Typically, these spurious responses show reaction times of under one second. To prevent such participants from affecting the analyzed data, the mean reaction time on critical items was calculated by participant and those participants with mean reaction times under 2000 ms were excluded from the analysis. This trimming process is relatively conservative insofar as it potentially leaves in other very fast (e.g., participant pressed a button by mistake) or very slow (e.g., participant stopped paying attention) responses. In Experiment 1, 3 participants were excluded from the analysis on the basis of their mean reaction time.

Among those participants who were included in the analysis, the mean completion time for the experiment, including the demographic survey and Worker ID entry, was 11 minutes and 28 seconds. Participants were paid USD 1.50 through AMT for completing the experiment.

3.1.4 Predictions

As stated above, the discussion for each of the experiments will focus on just two models of VPE interpretation: the strict syntactic account, termed the *identity* model, and the simple discourse account, termed the *discourse* account.

A strict identity account of VPE interpretation suggests that when possible, the meaning of an ellipsis site should always be calculated using only the exact information available in the antecedent utterance. In other words, the interpretation of the ellipsis site should be the same as if the VP from the antecedent utterance also appeared *in situ* in the ellipsis site. For the Experiment 1 sample item, this means that a VPE reply like *We can't* should be interpreted the same as *We can't buy candy bars* when the antecedent is unmodified, since that antecedent contains the VP *buy candy bars*, and the same as *We can't buy five candy bars* when the antecedent is modified, since that antecedent has the VP *buy five candy bars*.

In terms of the preferred interpretation, then, the identity account predicts that the polar interpretation should be preferred with an unmodified antecedent, since *We can't buy candy bars* entails that no candy bars can be purchased. By contrast, the scalar interpretation may be preferred with a modified antecedent. *We can't buy five candy bars* is compatible both with buying no candy bars and with buying some number of candy bars less than five. If the “fewer than five” interpretation for such an utterance is dominant, then the scalar interpretation will be preferred.

More critically for comparing the two models of interpretation, the strict identity account also predicts that in the presence of an unmodified or modified antecedent, the manipulation in the comic strip context should have no effect on the preference for the polar versus the scalar interpretation. This is because a strict formulation of this account holds that VPE is always resolved using only information from the linguistic antecedent when possible. Since both the unmodified and modified antecedents are sufficient for resolution under linguistic identity, this account predicts that information in the broader discourse context should be ignored in these conditions.

The identity account does not directly make a prediction about the results in the exophoric conditions. Antecedentless ellipsis is “defective” from the perspective of the strict identity account, as there is by definition no linguistic antecedent to use for resolution under identity. As discussed in the previous chapter, a strict identity model must include an “escape hatch” to account for the interpretation of antecedentless ellipsis, such as Merchant’s 2004 proposal that listeners infer that the elided construction is *do that* in the case of antecedentless VPE. In general, such proposals likely lead to the prediction that the interpretation in the exophoric conditions will be isomorphic to the general discourse availability of the competing interpretations.

In contrast to the identity model, a simple discourse model suggests that there should be a straightforward mapping between the discourse availability of the competing interpretations and the distribution of interpretations for a VPE construction. Critically for Experiment 1, this means that the discourse account is compatible with an effect of the comic strip context even in the presence of a linguistic antecedent, unlike the identity account. This is because a manipulation in the broader discourse context could conceivably modulate the discourse availability of one or more potential VPE interpretations even if the linguistic antecedent is the dominant factor in determining discourse salience. However, without data assessing the discourse availability of the competing interpretations in each discourse context prior to the VPE utterance, it is impossible to directly assess the validity of the simple discourse model, since a comparison of the discourse availability of the competing interpretations and their availability under ellipsis is not possible. This is addressed below in Experiment 3.

Thus, the analysis presented below focuses primarily on detecting an effect of changing the comic strip context on the strength of the two types of VPE interpretation in the presence of a viable linguistic antecedent (unmodified or modified). Such an effect is predicted to be possible by the discourse account, but not by the strict identity account.

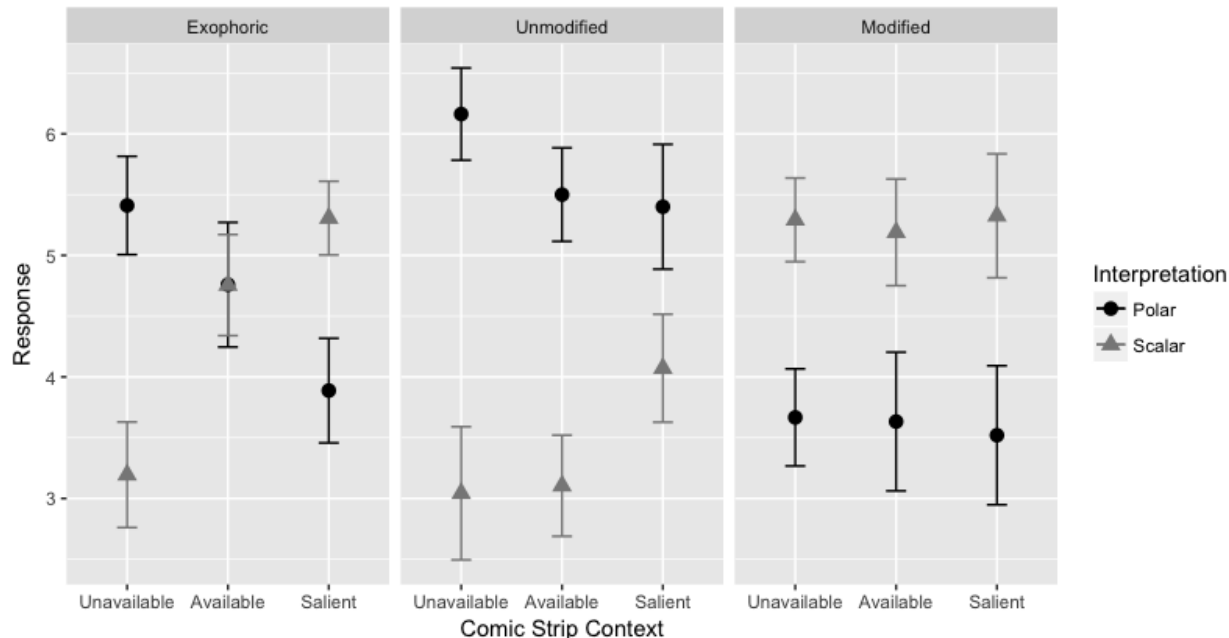


Figure 3.2: Experiment 1 results. Frame split: Antecedent. Error bars: 95% confidence interval.

3.1.5 Results

The aggregate results by condition are shown in Figure 3.2. Additionally, results for each of the 6 critical experimental items are shown in Appendix A.

A visual inspection of the plot suggests that there are differential effects of the comic strip context manipulation depending on the antecedent. For the exophoric conditions, the preferred interpretation appears largely to be a function of the intended salience manipulation. The polar interpretation appears to be preferred in the unavailable context, where there is no information about the critical numeral in the comic strip context. As information about the critical numeral becomes more salient, the polar interpretation degrades and the scalar interpretation improves until the scalar interpretation is preferred in the salient context.

For the unmodified antecedent conditions, the polar interpretation appears to be preferred regardless of the comic strip context. This is not surprising, as the polar interpretation would arise if the VPE reply were interpreted with the exact antecedent VP *in situ*, and this VP lacks a numeral. However, the plot suggests that there is an effect of the comic strip context

manipulation, although the magnitude of this effect is not enough to change the preferred interpretation. The ratings for the polar interpretation appear to be somewhat higher in the unavailable condition compared to the available and salient conditions, suggesting a slight degradation of this interpretation as numeral information becomes more salient. Likewise, the scalar interpretation appears to have received slightly better ratings with a salient context than with an unavailable or available context, suggesting that that interpretation improved with increasing number salience.

Finally, with a modified antecedent, the scalar interpretation appears to be preferred across the board, with no clear effect of the comic strip context manipulation on the ratings for either interpretation. Again, this is not surprising considering the antecedent. Since the antecedent contains a numeral, an identical interpretation of the VPE reply, which includes a numeral under negation, intuitively makes the “not so many” interpretation quite available.

3.1.6 Analysis

The response data were fit to a linear mixed effects regression model with a three-way interaction of comic strip context, antecedent, and interpretation prompt, plus component two-way interactions and main effects of each, along with random intercepts for participant and item². There was a significant three-way interaction of comic strip context, antecedent, and interpretation prompt ($p < .001$), indicating that the effect of the comic strip context manipulation on the ratings for each interpretation differs by antecedent.

On the basis of this three-way interaction, the data were split up by antecedent type for further analysis. The data for each antecedent were fit to a linear mixed effects regression model with a two-way interaction of comic strip context and interpretation prompt, main effects of each, and random intercepts for item and participant.³

For the exophoric data, there was a significant interaction between comic strip context

2. **lmer** model specification: $\text{Response} \sim \text{Context} * \text{Antecedent} * \text{Interpretation} + (1 \mid \text{Participant}) + (1 \mid \text{Item})$.

3. **lmer** model specification: $\text{Response} \sim \text{Context} * \text{Interpretation} + (1 \mid \text{Participant}) + (1 \mid \text{Item})$.

and interpretation prompt ($p < .001$), indicating that the effect of the comic strip context manipulation differed by the interpretation under consideration. The main effects of comic strip context and interpretation prompt were not significant (p 's $> .1$).

For the unmodified antecedent conditions, the interaction between comic strip context and interpretation prompt was also significant ($p < .01$), again indicating a differential effect of the comic strip context manipulation depending on the interpretation under consideration. The main effect of interpretation prompt was significant ($p < .001$), suggesting that one of the interpretations was preferred across the board. The main effect of comic strip context was not significant ($p > .1$).

For the modified antecedent data, the interaction between comic strip context and interpretation prompt was not significant ($p > .8$). The main effect of interpretation prompt was significant ($p < .001$), indicating that one interpretation received higher ratings across the board. The main effect of comic strip context was not significant ($p > .9$).

To further explore the effect of the comic strip context manipulation on the competing interpretations, paired comparisons were carried out using estimated marginal means. For the exophoric and unmodified conditions, comparisons were made between comic strip contexts within interpretations, and between interpretations within comic strip contexts, on the basis of the significant interaction between these factors. The estimated marginal means were calculated using the linear mixed effects regression models described above.

For the exophoric data, the ratings for the polar interpretation were not significantly different between the unavailable and available contexts ($p > .1$), but were significantly different between the unavailable and salient ($p < .001$) and available and salient contexts ($p < .05$). The ratings for the scalar interpretation were significantly different between the unavailable and available and the unavailable and salient contexts (p 's $< .001$), but not between the available and salient contexts ($p > .2$). The ratings for the two interpretations were significantly different in the unavailable and salient contexts (p 's $< .001$), but not the available context ($p > .9$).

For the unmodified conditions, the ratings of the polar interpretation were not significantly different in any pair of conditions ($p > .1$). The ratings for the scalar interpretation were not significantly different between the unavailable and available contexts ($p > .9$), but were significantly different between the unavailable and salient and between the available and salient contexts (p 's $< .01$). The ratings for the two interpretations were significantly different in all three context conditions (p 's $< .001$).

Paired comparisons were not carried out for the modified antecedent data since only the main effect of interpretation prompt was significant; inspection of Figure 3.2 suggests that the scalar interpretation was uniformly preferred.

3.1.7 Discussion

In general, the regression analysis and paired comparisons confirmed the preliminary conclusions from the visual inspection of Figure 3.2. For the exophoric conditions, the strength of the two VPE interpretations was largely a function of the salience of the critical number information in the comic strip context. With an unavailable context, where the number information is not salient, the polar interpretation was preferred. With increasing salience in the comic strip context, the polar interpretation degraded and the scalar interpretation improved until the scalar interpretation was preferred in the salient context. This indicates that participants were sensitive to the intended manipulation and confirms that antecedentless VPE resolution proceeds largely along the lines of discourse salience.

For the unmodified antecedent conditions, the polar interpretation was preferred in the presence of all three comic strip contexts. This makes sense according to both the identity and discourse accounts. Since the antecedent VP does not contain a numeral, the identity account predicts that the VPE reply should not be interpreted as numeral-modified. Likewise, the discourse account predicts that number information should not figure prominently in the interpretation of the VPE reply, since a number-unmodified proposition is likely most salient after the unmodified antecedent is uttered.

Critically, however, there is still an effect of the comic strip context manipulation on the interpretation of the VPE reply even in the presence of a viable linguistic antecedent. While the apparent trend of degradation for the polar interpretation between the unavailable and available contexts did not attain significance, the scalar interpretation was rated significantly better in the salient context than in the unavailable or available contexts. This indicates that salient information in the broader discourse context can affect the interpretation of VPE in ways that are not directly supported by the content of a recoverable linguistic antecedent.

For the modified antecedent conditions, the scalar interpretation was preferred in all three comic strip context conditions, with no apparent improvement or degradation for either interpretation in response to the manipulation of the salience of number information. Again, the preference for the scalar interpretation makes sense according to both interpretation models. Since the antecedent contains a numeral, the identity account predicts that the ellipsis site should be interpreted as containing a numeral, making the scalar interpretation intuitively very available. The discourse account predicts that a number-modified proposition should be highly salient when a numeral is uttered in the antecedent, which should also make the scalar interpretation for VPE highly available.

Considered together, the results provide preliminary evidence against a strict identity account of VPE interpretation in complex discourse contexts. This conclusion is driven largely by the findings in the unmodified antecedent conditions. The identity model predicts that VPE in these conditions should uniformly be interpreted as identical to the antecedent VP – that is, number-unmodified. However, the distribution of interpretations available under ellipsis was actually sensitive to the manipulation of the salience of number information in the broader linguistic context, with increasing salience improving ratings for the scalar interpretation and promoting a numeric trend toward degradation for the polar interpretation.

Interestingly, the effect of the comic strip context manipulation observed for the unmodified conditions does not hold for the modified antecedent conditions, where the ratings for the two interpretations were virtually identical in all three context conditions. This may

point to an asymmetry in the possible effects driven by information in a broader discourse context compared to a linguistic antecedent. The unmodified antecedent conditions suggest that salient information in the broader discourse context can augment the information that is considered to be at-issue relative to the linguistic antecedent, since number information was considered to be somewhat more at-issue when it was salient in the comic strip context. By contrast, when a numeral was already overtly used in the antecedent, there is no evidence that the salience of broader contextual information affects interpretation, either by further promoting the use of number information when also salient in the broader context, or by slightly inhibiting its consideration when not salient.

In contrast to the identity account, the findings from Experiment 1 do not directly test the predictions of the discourse account. The discourse model is strictly compatible with the finding that information from the broader linguistic context affects VPE interpretation even in the presence of a linguistic antecedent that could be used for resolution under identity. However, a strict formulation of the discourse account predicts that there should be a one-to-one mapping between the interpretations that arise under VPE and the discourse salience of the competing propositions in the broad discourse context. Since Experiment 1 did not assess the discourse availability of the competing interpretations in the absence of ellipsis, it cannot yet be determined whether this prediction of the strict discourse account is borne out. This is addressed below in Experiment 3.

3.2 Experiment 2: Ellipsis-specific context effects

Experiment 1 showed preliminary evidence that a strict identity model of VPE interpretation cannot fully account for the distribution of interpretations that arise in complex discourse contexts. The identity account holds that an ellipsis site should be interpreted only using the content of the linguistic antecedent VP, when one is available. However, the Experiment 1 results showed that variations in the information available in the broader discourse context can affect the distribution of interpretations even in the presence of the same viable linguistic

antecedent.

Recall that the critical contextual effect found in Experiment 1 was that with an unmodified antecedent, salient number information in the broader discourse context could boost the number-modified scalar interpretation to a significant degree compared to contexts where number information was not salient in the context. One explanation for the effect found in Experiment 1 that maintains a strict identity model of VPE interpretation is that it is driven by an effect of the contextual manipulation on utterance interpretation generally, rather than on elliptical replies in particular. Thus, since the strict identity model predicts that the example VPE reply *We can't* with the unmodified antecedent *I want to buy candy bars* should be interpreted as meaning the same thing as *We can't buy candy bars*, this alternative explanation predicts that the interpretation of the full-VP sentence *We can't buy candy bars* is itself sensitive to the manipulation of the salience of number information in the broader linguistic context in a way that is isomorphic to the pattern of results from Experiment 1.

The goal of Experiment 2 is to determine whether this alternative explanation can sufficiently explain the pattern of results found in Experiment 1, thus maintaining the viability of the strict identity account of VPE interpretation. This is investigated by replacing the VPE reply in the unmodified and modified antecedent conditions with full-VP counterparts. In particular, the reply sentences contain the same VP as the antecedent sentence in each case, excluding a numeral in the case of the unmodified antecedent conditions and including one in the modified conditions. Thus, the experiment elicits ratings of interpretations for the exact structures predicted to be under consideration in Experiment 1 according to the strict identity account, meaning that a comparison of the patterns of results for the two experiments should indicate whether the pattern for Experiment 1 was driven by a contextual effect on interpretation availability that affects full-VP as well as elliptical sentences.

3.2.1 *Design, materials, and procedure*

The design, materials, and procedure for Experiment 2 were largely identical to those for Experiment 1. The only exception is that in the unmodified and modified antecedent conditions, the VPE reply was replaced with a sentence containing a fully specified VP, namely the same VP that appeared in the antecedent utterance. Thus, in the sample scenario, the VPE reply utterance *We can't* was replaced with *We can't buy candy bars* in the unmodified antecedent condition, and with *We can't buy five candy bars* in the modified antecedent condition, as shown in Table 3.2. Since there is no antecedent VP in the exophoric conditions, the reply utterance was left as a VPE construction, and these conditions merely replicate the same conditions from Experiment 1.

The practice items and fillers were adjusted to match the task for the critical items such that items with a linguistic antecedent featured a reply with an identical VP, whereas items that had an antecedentless VPE reply in Experiment 1 were identical in Experiment 2. The full specifications for all items in the experiment are shown in Appendix B.

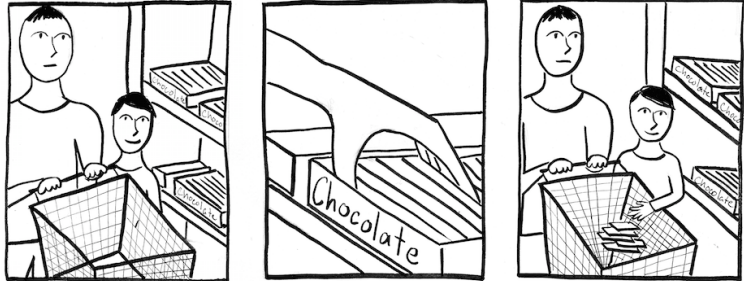
3.2.2 *Participants*

204 participants took part in Experiment 2. 110 were female and 94 were male. The mean age of the participants was 34.3 years. The data from 6 participants was excluded from analysis because they failed to identify themselves as native English speakers in the demographic survey. The data from a further 3 participants was excluded because their mean reaction time for the critical items and fillers was under 2000 ms. Among the participants whose data was analyzed, the mean completion time for the experiment, including the demographic survey, instructions, and AMT cross-referencing screen, was 9 minutes, 32 seconds. Participants were paid USD 1.50 for completing the experiment.

Unavailable comic



Available comic



Salient comic



Exophoric antecedent
& reply

[no utterance shown]
Father: We can't.

Unmodified antecedent
& reply

Son: I want to buy candy bars!
Father: We can't buy candy bars.

Modified antecedent
& reply

Son: I want to buy five candy bars!
Father: We can't buy five candy bars.

Polar prompt

On a scale from 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the father meant: We can't buy any candy bars today.

Scalar prompt

On a scale from 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the father meant: We can't buy five candy bars today, but maybe we could buy fewer.

Table 3.2: Sample stimuli for Experiment 2

3.2.3 Predictions

Experiment 2 tests whether the critical contextual effect observed in the unmodified-antecedent conditions of Experiment 1 is driven by a general process that also affects sentences with fully specified VPs, in which case the strict identity account of VPE interpretation is still viable, or whether the effect is unique to ellipsis, which would suggest that the strict identity model does not adequately capture the facts of VPE interpretation in context.

If the critical Experiment 1 results are due to a general interpretive effect of salient information in the broader discourse context, regardless of the presence of ellipsis, then a similar pattern of results should be seen in Experiment 2. In particular, this hypothesis predicts that the interpretation of the numeral-unmodified full-VP sentence *We can't buy candy bars* should show an effect of the manipulation of the salience of number information in the broader context, with the strength of the scalar interpretation increasing as number information is made more salient.

By contrast, if the Experiment 1 effect is specific to ellipsis, then there should not be an effect of the salience manipulation in the broader context on the interpretation of the full-VP replies in the unmodified and modified conditions. Under this hypothesis, the interpretation of full-VP utterances like *We can't buy candy bars* is fixed, and the Experiment 1 effect is due to an ellipsis-specific interpretation mechanism (for instance, the discourse account of interpretation).

3.2.4 Results

The results of Experiment 2 are shown in Figure 3.3. In addition, by-item results are shown in Appendix B.

A visual inspection of Figure 3.3 suggests that for the exophoric conditions, the results of Experiment 1 were largely replicated. While the polar interpretation was preferred with an unavailable context, the strength of the polar interpretation decreased and the strength of the scalar interpretation decreased as number information became more salient in the

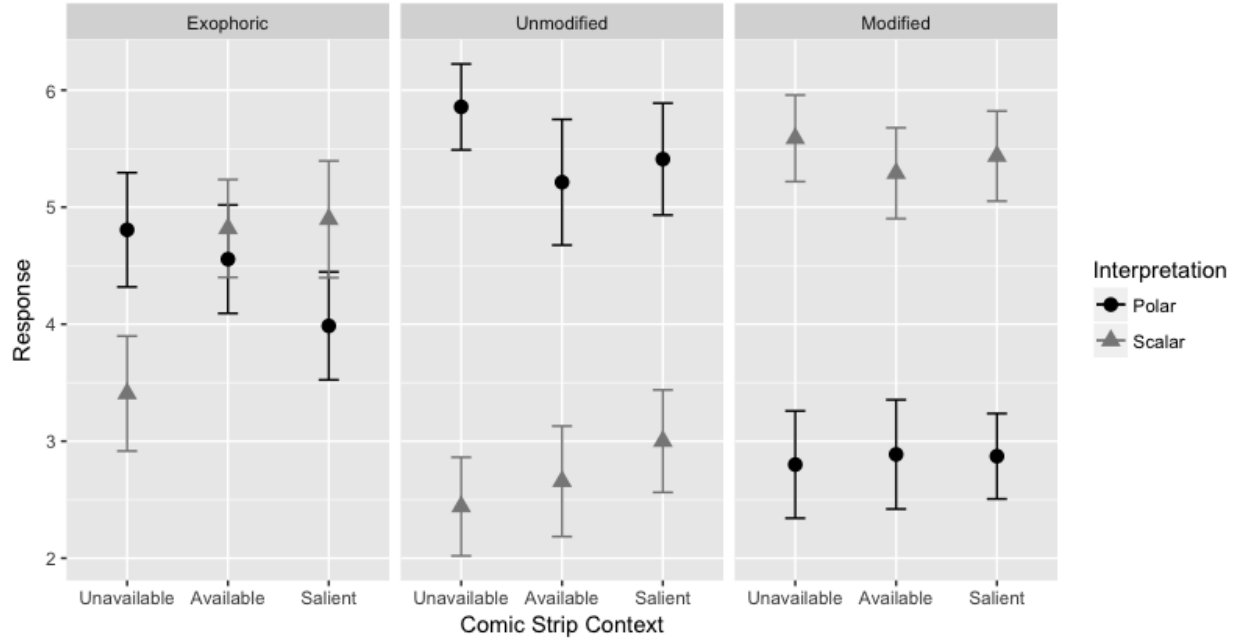


Figure 3.3: Experiment 2 results. Frame split: Antecedent. Error bars: 95% confidence interval.

broader discourse context, although the magnitude of the effect appears to be smaller than in Experiment 1.

For the unmodified antecedent and reply, there does not appear to be a strong effect of the contextual manipulation on the strength of the two interpretations. For the polar interpretation, there appears to be a trend that the ratings with an available and salient context are lower than those with an unavailable context, but it is not clear that this difference is significant. Likewise, for the scalar interpretation, there is an apparent trend toward increasing ratings with increasing number salience in the broader context, but the magnitude of this effect appears smaller than in Experiment 1.

For the modified conditions, there again appears to be no effect of the contextual manipulation. The scalar interpretation is preferred across the board, and the ratings for both interpretations neither increase nor decrease with increasing number salience in the comic strip context; inspection of Figure 3.3 indicates that the scalar interpretation was uniformly preferred.

3.2.5 Analysis

The data were fit to a linear mixed effects regression model with a three-way interaction of comic strip context, antecedent/reply combination, and interpretation prompt, plus two-way interactions and main effects of each of these factors, and random intercepts for item and participant. There was a significant three-way interaction in this model ($p < .001$), suggesting that the combined effect of interpretation prompt and comic strip context was different for each antecedent/reply combination.

On the basis of the significant three-way interaction, further analysis was carried out separately for each antecedent/reply combination. For each combination, the response data were fit to a linear mixed effects regression model with an interaction between comic strip context and interpretation prompt, main effects of each, and random intercepts for item and participant.

For the exophoric model, there was a significant interaction of comic strip context and interpretation prompt ($p < .001$), indicating that the effect of the comic strip context manipulation was different for the two interpretation types. The main effect of comic strip context was significant ($p < .05$), while the main effect of interpretation prompt was not ($p > .6$).

For the unmodified antecedent/reply conditions, there was a marginal interaction of comic strip context and interpretation prompt ($p < .1$). The main effect of interpretation prompt was significant ($p < .001$), while the main effect of comic strip context was not ($p > .4$).

For the modified antecedent/reply conditions, the interaction between comic strip context and interpretation prompt was not significant ($p > .6$). The main effect of interpretation prompt was significant ($p < .001$), while the main effect of comic strip context was not ($p > .8$).

Paired comparisons were carried out using the estimated marginal means for the exophoric and unmodified models to further investigate the significant or marginal interaction terms. For the exophoric model, the ratings of the polar interpretation were not significantly different between the unavailable and available ($p > .7$) or between the available and salient contexts ($p > .1$), but the difference between the unavailable and salient context was

significant ($p < .05$). For the scalar interpretation, the differences between then unavailable and available and between the unavailable and salient contexts were significant (p 's $< .001$), while the difference between the available and salient context was not significant ($p > .9$). In the exophoric conditions, the polar and scalar interpretations received significantly different ratings with an unavailable ($p < .001$) and salient context ($p < .01$), but not with an available context ($p > .4$).

For the unmodified antecedent/reply conditions, the ratings for the polar interpretation were not significantly different for any pair of comic strip contexts (unavailable-available $p > .1$, unavailable-salient $p > .3$, available-salient $p > .8$). Likewise, the ratings for the scalar interpretation were not significantly different for any context pair (unavailable-available $p > .7$, unavailable-salient $p > .1$, available-salient $p > .5$). The polar and scalar interpretations received significantly different ratings in all three comic strip context conditions (p 's $< .001$).

Paired comparisons were not carried out for the modified antecedent/reply conditions because only the main effect of interpretation was significant.

3.2.6 *Discussion*

The results of Experiment 2 suggest that the critical context effect observed in Experiment 1 was not completely driven by an ellipsis-independent effect of salient information in the broader context on utterance interpretation. The results for the exophoric and modified conditions in Experiment 2 largely mirrored those from Experiment 1. However, the results for the unmodified conditions suggest that the effect from Experiment 1 is at least partially specific to ellipsis.

Turning to the unmodified conditions, in Experiment 1, the ratings of the scalar interpretation with an unmodified antecedent were significantly higher in the presence of a salient context than in the presence of an unavailable or available context. In Experiment 2, there was a numeric trend toward improved ratings for the scalar interpretation with increasing number salience in the comic strip context. However, there was only a marginally significant

interaction between the comic strip context and the interpretation prompt in Experiment 2, and paired comparisons indicated that the scalar interpretation did not receive significantly different ratings in any of the three context conditions. As in Experiment 1, the ratings for the polar interpretation were largely unaffected by the comic strip context manipulation in Experiment 2.

The Experiment 2 results indicate that the interpretation of numeral-unmodified, full-VP sentences like *We can't buy candy bars* is largely unaffected by variations in the information that is salient in the broader discourse context. This suggests that the effect observed in Experiment 1, where salient number information in the broader context promoted a number-modified interpretation of a VPE reply utterance beyond its baseline level, is at least in part unique to ellipsis. While there was a numeric trend toward improved ratings of this interpretation for full-VP sentences, this effect was not significant and its magnitude was less than what was seen in Experiment 1. However, on the basis of this numeric trend, it may be prudent to consider that the critical effect in Experiment 1 was driven partly by a non-ellipsis-specific effect of salient number information and partly by an ellipsis-specific sensitivity to information available in the broader discourse context.

Thus, the Experiment 2 results failed to detect evidence that the contextual effect observed in the unmodified antecedent conditions of Experiment 1 was due to ellipsis-independent changes in interpretation preferences in response to changes in the information salient in the broader discourse context. Since the distribution of interpretations available for the VPE reply utterances in Experiment 1 cannot be reduced to the interpretations that are available for the corresponding full-VP utterances containing the antecedent VP *in situ*, the two experiments suggest that a strict linguistic identity model cannot account for the facts of VPE interpretation in complex discourse contexts. The next experiment investigates the viability of the competing hypothesis, the simple discourse account of VPE interpretation.

3.3 Experiment 3: Discourse availability of VPE interpretations

Experiment 1 showed that the distribution of interpretations available for an ellipsis site can depend at least in part on the information that is salient in the broader discourse context, even when there is a linguistic antecedent that could be used to resolve the ellipsis strictly along identity lines. Experiment 2 investigated the possibility that this effect was driven by changes in the interpretations available for utterances with fully realized VPs in response to manipulations in the information available in the broader context – in essence, that the interpretations observed in Experiment 1 were actually the same as the interpretations for the full-VP utterances corresponding to the resolution of the ellipsis site under strict identity. The results indicated that the critical effect from Experiment 1 was due at least in part to the presence of ellipsis. Thus, the two experiments together are problematic for the strict identity account of VPE interpretation, which predicts that the interpretation of an ellipsis site should not change in response to manipulations in the broader discourse context given a constant linguistic antecedent that could be used for resolution under identity.

The goal of the current experiment is to open investigation of the competing hypothesis, the simple discourse model of VPE interpretation. According to this model, VPE is a form of discourse reference that is resolved with reference to the information that is salient in the discourse, including not only a linguistic antecedent if one is available, but also the information in the broader discourse context. Under a simple discourse model, competing interpretations for an ellipsis site should be considered in proportion to their availability in the discourse at large – in other words, there should be a roughly one-to-one mapping between how salient discourse participants find certain information in the context and how likely they are to say a VPE utterance targets that information.

Experiment 3 investigates this proposal by eliciting participants’ assessments of how coherent the Experiment 1 interpretation prompts were in each of the discourse contexts (comic strip context and antecedent combination). By hypothesis, if a number-unmodified proposition (e.g., the son wanting to buy candy bars, but not a particular number) is highly

at-issue in the discourse, then the polar interpretation prompt (where the father rules out the possibility of buy candy bars of any number) should be rated as highly coherent. If a number-modified proposition (e.g., the son wanting to buy specifically five candy bars) is at-issue, then the scalar prompt (where the father objects to the specific number of candy bars) should be rated as highly coherent. Of course, it is not intuitively incoherent for the father to object to buying candy bars altogether when the son clearly has a specific number in mind. The goal of the experiment is thus primarily to establish whether there is a correspondence between the more coherent reply in each context and the preferred interpretation under VPE rather than to establish a direct numeric mapping between the ratings in the two experiments. (The next chapter features a revised experimental paradigm that can more directly test this numeric mapping.)

3.3.1 Design, materials, and procedure

The design, materials, and procedure for Experiment 3 were largely identical to those for Experiment 1. However, instead of a VPE reply utterance followed by either a polar or scalar interpretation prompt, the reply utterance was changed to either the exact polar or scalar interpretation prompt from Experiment 1. Thus, the experiment had a 3 (comic strip context) x 3 (antecedent) x 2 (reply) design. After the reply utterance, participants always saw a uniform prompt asking them to rate the coherence of the reply utterance considering the information available in both the comic strip context and the antecedent utterance, if there was one. The practice items and fillers were adjusted in the same way. Materials for the sample scenario are shown in Table 3.3, and the full set of materials is shown in Appendix C.

3.3.2 Participants

196 participants took part in the experiment. 108 were female, and 88 were male. The mean age of the participants was 35.1 years. The data from 4 participants was excluded

Unavailable comic



Available comic



Salient comic



Exophoric antecedent	[no utterance shown]
Unmodified antecedent	<i>Son:</i> I want to buy candy bars!
Modified antecedent	<i>Son:</i> I want to buy five candy bars!
Polar reply	<i>Father:</i> We can't buy any candy bars today.
Scalar reply	<i>Father:</i> We can't buy five candy bars today, but maybe we could get fewer.
Prompt:	On a scale from 1 to 7, where 1 is the least coherent and 7 is the most coherent, how coherent is the father's statement considering what came before?

Table 3.3: Sample stimuli for Experiment 3

from the analysis because they failed to identify themselves as native speakers of English in the demographic survey. The data from a further 6 participants was excluded because they had a mean reaction time of less than 2000 ms on critical items and fillers. Among participants whose data was analyzed, the mean time to complete the experiment, including the demographic survey, instructions, and AMT information page, was 10 minutes, 36 seconds. Participants were paid USD 1.50 through AMT for completing the experiment.

3.3.3 Predictions

The simple discourse account of VPE interpretation predicts that there should be a straightforward mapping between the discourse coherence of the polar and scalar replies and their availability under ellipsis as measured in Experiment 1. This is because the discourse account holds that VPE resolution is strictly a matter of discourse reference, with the proposition that is the most available in the discourse context also being the most likely target for resolution. The setup of the two experiments does not allow for a direct numeric comparison, as all of the replies in Experiment 3 are intuitively fairly coherent, whereas some of the ratings in Experiment 1 were quite poor. However, the qualitative preference for one reply over the other in each of the conditions can easily be compared to the preference for one interpretation under ellipsis to arrive at a preliminary assessment of the strength of the strict discourse model. The discourse account predicts that the preferred reply in Experiment 3 should also be the preferred ellipsis interpretation in Experiment 1.

If the preferred reply does not match the preferred interpretation outlined above, this constitutes evidence against a simple discourse model of VPE interpretation. One pattern of results that seems particularly plausible is a shift in interpretation preferences under ellipsis toward the antecedent-identical interpretation. For instance, in an unmodified condition, it may be the case that in Experiment 3, the two replies are rated as equally coherent, whereas in Experiment 1, the polar interpretation always received significantly higher ratings than the scalar interpretation. This would indicate that although information in the broader discourse

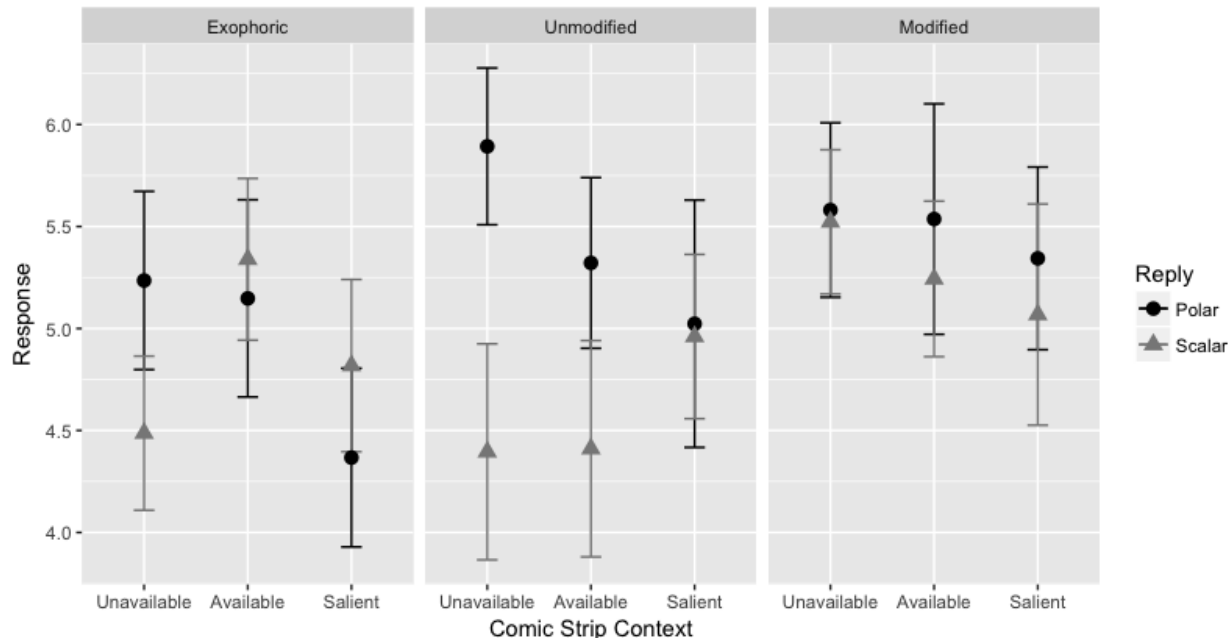


Figure 3.4: Experiment 3 results. Frame split: Antecedent. Error bars: 95% confidence interval.

context did affect VPE interpretation in Experiment 1, interpretation is still constrained by considerations of linguistic identity, since the interpretation corresponding to the linguistic antecedent would be preferred in excess of any advantage over the competing interpretation in discourse coherence. In light of the effect already found in Experiment 1, such a finding would point to a more nuanced account of VPE interpretation, where both identity considerations and information from the broader context are relevant.

3.3.4 Results

The results from Experiment 3 are shown in Figure 3.4. In addition, by-item results plots are shown in Appendix C.

A visual inspection of the plot again reveals different effects depending on the antecedent. For the exophoric conditions, there appears to be a small drop in the coherence of the polar reply in the salient context compared to the unavailable and available contexts. The scalar reply appears to be more coherent in an available context than in an unavailable one, but the

ratings decrease again in the salient context. This pattern is somewhat surprising, as both VPE interpretation models resort to discourse availability to explain VPE interpretation without a linguistic antecedent, and the results of both Experiments 1 and 2 showed a more direct mapping between number salience in the comic strip context and the preferred interpretation.

For the unmodified antecedent conditions, there is a clear coherence preference for the polar reply over the scalar reply with an unavailable context. As number information becomes more salient in the comic strip context, the ratings for the polar reply decrease. The ratings for the scalar reply stay roughly the same with an available context, but jump somewhat higher with a salient context, similar to the scalar interpretation in Experiment 1. With a salient context, the preference for the polar reply has eroded and the two replies are rated as roughly equally coherent.

For the modified antecedent conditions, there is not a clear preference for either reply in any of the three comic strip context conditions. The ratings for both replies, and in particular the scalar reply, appear to decrease with increasing number salience in the comic strip context, but the effect does not appear to be significant.

3.3.5 Analysis

The ratings were fit to a linear mixed effects regression model with a three-way interaction of comic strip context, antecedent, and reply utterance, two-way interactions for each pair of factors, main effects for each factor, and random intercepts for participant and item. This model had a marginal three-way interaction ($p < .1$).

To further explore this interaction, the data were analyzed separately by antecedent. For each antecedent level, the data were fit to a linear mixed effects regression model with an interaction between comic strip context and reply, main effects of each, and random intercepts for participant and item.

For the exophoric model, there was a significant interaction of comic strip context and

reply ($p < .05$) and a significant main effect of comic strip context ($p < .05$), while the effect of reply utterance was not significant ($p > .6$).

For the unmodified conditions, there was a significant interaction of comic strip context and reply ($p < .05$) and a significant main effect of reply utterance ($p < .001$), while the effect of comic strip context was not significant ($p > .3$).

For the modified conditions, there was not a significant interaction ($p > .2$), nor were the main effects of reply ($p > .3$) or comic strip context ($p > .2$) significant.

For the exophoric and unmodified conditions, to further explore the significant interactions, estimated marginal means were used with the regression models to make paired comparisons. For the polar reply with no antecedent, the ratings with an unavailable and available context were not significantly different ($p > .9$), but the ratings with a salient context were significantly different from those with both an unavailable and an available context (p 's $< .05$). For the scalar reply, the unavailable and available ratings were significantly different ($p < .05$), while the unavailable and salient ($p > .5$) and the available and salient ($p > .2$) ratings were not significantly different. The ratings for the two replies were significantly different with an unavailable context ($p < .05$), but not with an available ($p > .7$) or a salient ($p > .1$) context.

For the unmodified antecedent conditions, the polar reply had significantly different ratings in the unavailable and salient conditions ($p < .05$), but not in the unavailable and available conditions ($p > .1$) or the available and salient conditions ($p > .6$). The ratings for the scalar interpretation were not significantly different between any pair of context conditions (unavailable-available $p > .9$, unavailable-salient $p > .3$, available-salient $p > .1$). The polar and scalar replies received significantly different ratings in the unavailable ($p < .001$) and available ($p < .01$) contexts, but not in the salient context ($p > .9$).

Paired comparisons were not carried out for the modified antecedent conditions because there was not a significant effect of any of the experimental manipulations on the ratings.

3.3.6 Discussion

The results of Experiment 3 are largely problematic for the discourse account of VPE interpretation because the more coherent reply utterance does not reliably predict the preferred VPE interpretation from Experiment 1. In the exophoric conditions, there was a trend toward decreasing ratings for the polar reply and increasing ratings for the scalar reply as number information became more salient in the comic strip context, with the exception of the surprisingly low rating for the scalar reply with a salient context. These findings are largely consistent with the Experiment 1 results, where the ratings for the polar interpretation decreased and the ratings for the scalar interpretation increased as number information became more salient. While the coherence rating for the scalar reply with a salient context is quite low in Experiment 3, there is still a numeric trend favoring it over the polar reply, and the scalar interpretation is significantly preferred in Experiment 1. The low rating of the scalar interpretation in the salient condition may be worth considering in the future, but will not be discussed more here.

With an unmodified antecedent, there was a trend toward decreasing ratings for the polar reply with increasing contextual number salience, as well as a non-significant boost for the scalar reply in the salient context only. These findings largely mirror the pattern of results in Experiment 1, where there was a non-significant drop in the ratings for the polar interpretation between the unavailable and available contexts and a significant jump in the scalar ratings from the available to the salient context. These findings support the conclusion that the effects observed in Experiment 1 were indeed due to the manipulation in the salience of number information in the broader discourse context.

However, analysis of the participants' preferred replies by condition reveals a problem for the simple discourse account. While the polar reply is significantly preferred over the scalar reply with an unavailable context, this advantage completely erodes as number information becomes more salient in the comic strip context, and the two replies received almost identical ratings with a salient context. By contrast, although there was an effect of contextual number

information on the ratings in Experiment 1, the antecedent-identical polar interpretation was always categorically preferred over the scalar interpretation.

Similarly, the lack of an effect of the comic strip context manipulation in the modified antecedent conditions mirrors the overall pattern from Experiment 1, but analysis of the coherence ratings is problematic for the discourse account. Whereas participants were largely indifferent between the two replies in Experiment 3, in Experiment 1, the scalar interpretation is preferred in all context conditions.

Thus, there are several conditions in which Experiment 3 participants rated the polar and scalar replies as approximately equally coherent, yet either the polar or scalar interpretation of VPE was significantly preferred in Experiment 1. There was not a straightforward mapping between the discourse coherence of a number-unmodified versus number-modified proposition and the preferred VPE interpretation, which suggests that VPE interpretation in context is not simply a matter of accessing the most salient proposition in the discourse context. Therefore, a comparison of the results of Experiments 1 and 3 suggests that the simple discourse account, like the strict identity account, is an inadequate model of VPE interpretation in complex discourse contexts. The results of the experiments instead indicate that the most likely model of VPE interpretation is one that hybridizes identity and discourse considerations, a topic which is elaborated on in the general discussion below after a brief digression to investigate the interpretation of the verbal anaphor *do that*.

3.4 Experiment 4: Interpretation of *do that* in context

With minor adjustments, the experimental paradigm presented in this chapter can be used not only to investigate the interpretation of VPE in complex discourse contexts, but also the interpretation of other context-dependent verbal expressions like *do that*, *do this*, and *do it*.⁴ Recall that Hankamer and Sag (1976) classify VPE as a *surface anaphor* (or, in Sag and

4. The experiment presented in this section, and an earlier form of the analysis and discussion, was previously published in conference proceedings form as Luce et al. (2018).

Hankamer’s (1984) terms, *ellipsis*), meaning that it can felicitously be used under identity with a linguistic antecedent, but not when the information supporting its interpretation comes only from the nonlinguistic context, as demonstrated in their examples in (53)-(54).

(53) The children asked to be squirted with the hose, so they were ∅.

(54) [Hankamer points gun offstage and fires, whereupon a bloodcurdling female scream is heard.]

#I wonder who was ∅.

(Sag and Hankamer, 1984)

By contrast, anaphors like *do it* and *do that* are classified as *deep*, or *model-interpretive* (Sag and Hankamer, 1984) anaphors, and can receive their interpretation either from a linguistic antecedent or from the information that is salient in the broader discourse context, as in (55)-(56).

(55) The children asked to be squirted with the hose, so we did it.

(56) [Hankamer points gun offstage and fires, whereupon a bloodcurdling female scream is heard.]

Jorge, you shouldn’t have done it.

(Sag and Hankamer, 1984)

Just as the experimental paradigm presented above was used to probe whether verb phrase ellipsis is resolved primarily along linguistic identity (i.e., “surface”) or discourse (i.e., “deep”) lines, it seems worthwhile to explore the resolution of other verbal anaphors to determine how closely it adheres to the predictions of the relevant theoretical models.

If the interpretation of these verbal anaphors is “deep,” as suggested by Hankamer and Sag, then there should be clear effects of variations in the broader context on the interpretations they can receive. In the current experimental paradigm, this might be assessed by comparing the interpretation scores to the results for Experiment 3, which assessed the coherence of the meanings of various replies given the different possible pre-reply discourses.

Beyond representing a preliminary experimental investigation of this expanded range of verbal anaphors, extending the paradigm to explore additional constructions will help verify its validity as a tool for measuring the interpretation of VPE. The impressionistic difference in interpretive possibilities for VPE and “deep” verbal anaphors like *do that* is widely acknowledged and not in dispute; regardless of the actual mechanism of VPE interpretation, it is clear that it is more constrained by the form of its linguistic antecedent than are deep verbal anaphors. Thus, if the experimental paradigm presented above faithfully represents the interpretive possibilities for VPE, then applying the paradigm to a deep verbal anaphor should show a clear shift in results, namely in favor of a greater influence on interpretation of information from the broader discourse context as opposed to the linguistic antecedent.

Experiment 4 applies the experimental paradigm from Experiment 1 to the deep verbal anaphor *do that*. The results indicate that *do that* is indeed more sensitive to information available in the broader discourse context than VPE. The results support the validity of the experimental paradigm as a means to assess the strength of competing interpretations of verbal anaphora and provide additional empirical support for the surface/deep divide reported in the literature. Unsurprisingly, comparison to the results of Experiment 3 suggests that the interpretation of *do that* may still be subject to linguistic identity considerations separate from the effect of the linguistic antecedent on discourse salience, indicating that *do that* introduces form-specific interpretation biases that go beyond a simple salience-driven interpretation mechanism. The discussion below covers the possible implications of the findings for VPE and *do that* interpretation for the nature of the surface-deep divide.

3.4.1 *Design, materials, and procedure*

The design, materials, and procedure for Experiment 4 were identical to those for Experiment 1, with the exception that the original VPE reply was modified to use the verbal anaphor *do that*. This was accomplished by adding *do that* in the position of the original ellipsis site in the reply utterances, immediately after the auxiliary verb. The practice items

and fillers were also adjusted accordingly. Materials for a sample item are shown in Table 3.4, while the full set of stimuli are shown in Appendix D.

3.4.2 *Participants*

102 participants took part in the experiment, of whom 54 were female and 48 were male. The mean age of the participants was 30.4 years. The data from 5 participants was excluded from analysis because they failed to self-report as native speakers of English in the demographic survey. One additional participant’s data was excluded because their mean reaction time was under 2000 ms, suggesting that they were completing the experiment as quickly as possible without fully attending to the experimental trials. Among those participants whose data was analyzed, the mean completion time for the AMT portion of the experiment was 10 minutes, 19 seconds, and participants were compensated with 1.50 USD for completing the experiment.

3.4.3 *Predictions*

It is widely assumed that *do that* more can more readily receive an interpretation from information available in the discourse context at large than can VPE (Hankamer and Sag, 1976; Tanenhaus and Carlson, 1990; Mauner et al., 1995). If this is correct, the experimental results should indicate that *do that* is more sensitive to the manipulation of the information available in the comic strip context than VPE was.

One way that this sensitivity might manifest is for relatively “weaker” information from the nonlinguistic context to influence interpretation of *do that* in a similar way to “strong” nonlinguistic information for VPE. Recall that number information becomes more salient as the nonlinguistic context changes from unavailable to available to salient. If the interpretation of *do that* is more sensitive to nonlinguistic information than is VPE, then the distribution of interpretations for *do that* in, for instance, the available context, where number information is weakly available in the nonlinguistic context, might be more similar to

Unavailable comic



Available comic



Salient comic



Exophoric antecedent	[no utterance shown]
Unmodified antecedent	<i>Son:</i> I want to buy candy bars!
Modified antecedent	<i>Son:</i> I want to buy five candy bars!
Reply	<i>Father:</i> We can't do that.
Polar prompt	On a scale from 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the father meant: We can't buy any candy bars today.
Scalar prompt	On a scale from 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the father meant: We can't buy five candy bars today, but maybe we could buy fewer.

Table 3.4: Sample stimuli for Experiment 4

the VPE distribution in a salient context, where number information is highly available in the broad discourse context, than to the VPE distribution in an available context.

Another area for exploration is the similarity of interpretations of *do that* to the distribution of salient propositions in the pre-anaphor discourse at large. The mapping can shed light on how information from linguistic antecedents and the broader context is recruited in “deep” anaphor interpretation. Since *do that* is “deeper” than VPE, impressionistically speaking, the mapping between the interpretation results and the Experiment 3 coherence results may be closer than the mapping between VPE interpretation and discourse coherence. On the other hand, a “deep” interpretation mechanism may still invoke complex interpretation heuristics, including strategies that privilege the linguistic antecedent beyond its contribution to discourse, so the mapping may not be perfect even for *do that*.

3.4.4 Results

The results for Experiment 4 are shown in Figure 3.5. A visual inspection of the plot suggests results similar to those for VPE from Experiment 1, but with a few notable differences that suggest a greater sensitivity to information from the broader discourse context on the part of *do that*, which will be discussed in greater detail in the discussion presented below.

As for VPE, in the exophoric conditions, there is a numerical preference for the polar interpretation in the unavailable context. As number information becomes more salient in the nonlinguistic context, the polar interpretation loses ground and the scalar interpretation gains ground, with the scalar interpretation preferred with a salient context. However, unlike VPE, where the polar and scalar interpretations received approximately equal ratings with an available context, there seems to be a clear preference for the scalar interpretation in this context for *do that*, suggesting that information in the nonlinguistic context can be less salient to achieve a similar shift in interpretation for *do that* compared to VPE.

For the unmodified-antecedent conditions, the polar interpretation is likewise preferred with an unavailable context. Again, scores for the scalar interpretation increase with in-

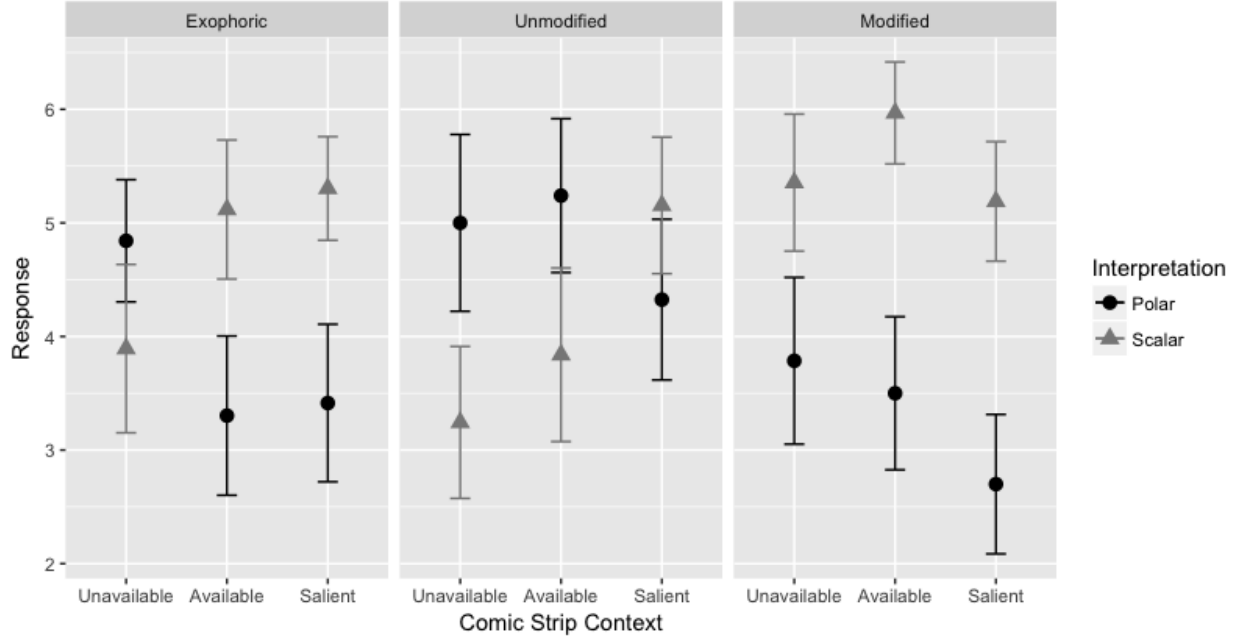


Figure 3.5: Experiment 4 results. Frame split: Antecedent. Error bars: 95% confidence interval.

creasing number information in the context, and ratings for the polar interpretation seem to decrease in the salient context relative to the unavailable and available contexts. In contrast to VPE, where the scalar interpretation improved in a salient context but did not overtake the polar interpretation, there is a numerical preference for the scalar interpretation of *do that* in this context.

For the modified-antecedent conditions, there is a global preference for the scalar interpretation. In contrast to VPE, there appears to be a trend of decreasing ratings for the polar interpretation as number information is made more salient in the nonlinguistic context. There is also a noticeable increase in ratings for the scalar interpretation with an available context relative to an unavailable or a salient context.

3.4.5 Analysis

A linear mixed-effects regression model was constructed with a three-way interaction between comic strip context, antecedent, and interpretation rated, plus the full component

effect structure (two-way interactions and component main effects), plus random effects for participant and item. The model showed a significant three-way interaction ($p < .01$).

This three-way interaction suggests that effect of the comic strip context and interpretation choice varies by antecedent, and on the basis of this interaction, further analysis was carried out separately for each antecedent. For each antecedent type, another linear mixed-effects regression model was constructed with an interaction between comic strip context and interpretation rated, main effects of each, and random effects for participant and item.

For the exophoric conditions, the model showed a significant interaction of comic strip context and interpretation ($p < .001$) and a significant main effect of interpretation rated ($p < .001$), while the effect of comic strip context was not significant ($p > .8$). The significant interaction indicates that the effect of the comic strip context manipulation on ratings depended on the interpretation being rated. Paired comparisons using estimated marginal means showed that the ratings for the polar interpretation were significantly different between the unavailable and available context ($p < .001$) and between the unavailable and salient context ($p < .01$), but not between the available and salient contexts ($p > .9$). The ratings for the scalar interpretation were different between the unavailable and available contexts ($p < .05$) and the unavailable and salient contexts ($p < .01$), but not between the available and salient contexts ($p > .9$). The two interpretations received significantly different ratings in all three contexts (unavailable $p < .05$, available $p < .001$, salient $p < .001$).

For the unmodified conditions, there was also a significant interaction of comic strip context and interpretation ($p < .001$) and a significant main effect of interpretation rated ($p < .01$), while the main effect of comic strip context was not significant ($p > .2$). Paired comparisons showed that the ratings of the polar interpretation were not significantly different in any pair of context conditions (unavailable-available $p > .9$, unavailable-salient $p > .2$, available-salient $p > .1$). The ratings for the scalar interpretation were significantly different between the unavailable and salient ($p < .001$) and the available and salient ($p < .05$) contexts, but not between the unavailable and available contexts ($p > .5$). The two interpretations had signif-

icantly different ratings in the unavailable ($p < .001$) and available ($p < .01$) contexts, while they were marginally different in the salient context ($p < .1$).

For the modified conditions, the interaction between comic strip context and interpretation rated was not significant ($p > .2$). However, the main effects of interpretation ($p < .001$) and comic strip context ($p < .05$) were both significant, indicating that both manipulations had an across-the-board effect on ratings that did not depend on the choice of level for the other factor. Paired comparisons were not carried out to further explore the effect of interpretation because there were only two levels. For comic strip context, paired comparisons were made using estimated marginal means without respect to the choice of interpretation. The ratings were not significantly different in the unavailable and available ($p > .8$) or the unavailable and salient conditions ($p > .1$), but there was a significant difference in the ratings in the available and salient conditions ($p < .05$).

3.4.6 Discussion

As mentioned above, the interpretation results for *do that* are reminiscent of those for VPE, but with some notable exceptions that suggest a greater sensitivity on the part of *do that* to salient information available in the broader (non-linguistic antecedent) discourse context.

In the exophoric conditions of Experiment 4, there was a significant interaction of comic strip context and interpretation rated, suggesting that the effect of the comic strip context was different for the two interpretations. The paired comparisons indicated that the polar interpretation was preferred with an unavailable context. However, as number information became more available in the comic strip context, the polar interpretation degraded and the scalar interpretation improved such that the scalar interpretation was preferred in both the available and salient contexts. As for VPE, this pattern of results is not surprising, as interpretation of *do that* must proceed entirely according to the features of the nonlinguistic context in the exophoric conditions, and so should closely track the nonlinguistic salience of number information.

In the unmodified-antecedent conditions, there was again a significant interaction of comic strip context and interpretation rated. The polar interpretation was again preferred over the scalar interpretation with an unavailable context, as well as in the available context. While a visual inspection of Figure 3.5 suggested a trend toward lower ratings for the polar interpretation as nonlinguistic number information became more salient, this trend did not attain significance. However, paired comparisons did indicate that the scalar interpretation was rated significantly higher in the salient context than in the unavailable or available contexts. The scalar interpretation was marginally preferred over the polar interpretation in the salient context. This general pattern is not surprising, as it echoes the unmodified-antecedent results for VPE, with a boost for the scalar interpretation only in the salient context.

In the modified-antecedent conditions, the most robust effect was that of the interpretation rated, with the scalar interpretation significantly preferred over the polar interpretation in all three contexts. This recalls the VPE results from Experiment 1, where the scalar interpretation was likewise preferred across the board with a modified linguistic antecedent. There was also a significant effect of comic strip context, although in the absence of a significant interaction, this effect suggests that the combined means for both the polar and scalar interpretations were affected by the context manipulation.

Paired comparisons and inspection of Figure 3.5 suggests that the effect of comic strip context is driven mainly by two data points. First, the low mean for the polar interpretation with a salient context likely lowered the overall mean for the salient conditions. This data point is not necessarily remarkable, as it is not surprising that the polar interpretation would degrade somewhat as nonlinguistic number information became more salient. Second, the high mean for the scalar interpretation in an available context raised the mean for the available-context conditions such that the overall means for the available and salient conditions were significantly different. It is not immediately clear why the scalar/available mean is higher than the corresponding unavailable and salient means. Further analysis or

experimentation may reveal whether this is simply due to experimental noise or is actually a robust effect driven by unforeseen complications in the interpretation of *do that*.

A major goal of Experiment 4 was to compare the results to those for both Experiment 1 (VPE interpretation) and Experiment 3 (discourse coherence of interpretations). This comparison and some possible implications for the theory of anaphora are discussed in the general discussion below.

3.5 General discussion

The main goal of this chapter was to explore the interpretation of verb phrase ellipsis in complex discourse contexts, meaning situations where the linguistic antecedent for an ellipsis site and the broader discourse context make different information available for interpretation. Complex discourse contexts are a potentially important testing ground for adjudicating between identity-driven and discourse-driven approaches to VPE interpretation. A strict identity-driven approach predicts that VPE interpretation should proceed entirely according to the linguistic antecedent when one is available, meaning manipulations in the information available in the broader context should be ignored when there is a linguistic antecedent. A simple discourse-driven account holds that VPE is interpreted according to the distribution of propositions that are available in the discourse at large, meaning the distribution of possible interpretations for VPE should be affected by both manipulations in the linguistic antecedent and in the broader context.

A secondary goal of the chapter was to demonstrate how the novel experimental paradigm could be extended to investigate additional context-dependent verbal constructions, such as the verbal anaphor *do that*. Since it is well established that *do that* is more sensitive to information in broad discourse contexts than VPE, this extension provides a useful opportunity to validate the experimental paradigm. At the same time, examining how multiple context-dependent constructions are interpreted in the same paradigm can provide important insights about the nature of context-dependent constructions more generally, in particular

with respect to whether their interpretive mechanisms can be given a uniform analysis or whether there are distinct classes of construction with separate mechanisms for interpretation. The results for VPE alone and the comparison between VPE and *do that* are discussed in turn below.

3.5.1 *VPE interpretation in context*

This chapter introduced a novel experimental paradigm used to assess the interpretation of VPE in complex discourse contexts. The paradigm critically manipulates the availability of number information about a particular referent in both the broad discourse context (comic strip context) preceding the ellipsis site, and in the ellipsis site’s linguistic antecedent. The experiments elicited participants’ assessments of the likelihood of two competing interpretations, one where number information was at issue and the other where it was not at issue. These responses were taken as indicative of whether the presence of number information in the broad discourse context or the linguistic antecedent, depending on the experimental condition, made participants more or less likely to consider number information in the interpretation of the ellipsis site.

Experiment 1 explored the interpretation of VPE in these discourse contexts. As expected, in the absence of a linguistic antecedent, whether number information was included in the interpretation was a function of how salient number information was in the discourse context at large. When a linguistic antecedent was present and included a numeral, participants preferred the interpretation taking number information into account across the board, which was judged to be consistent with both the identity and discourse models of interpretation. When there was a linguistic antecedent with no numeral, the number-unmodified interpretation was preferred across the board. However, the number-modified interpretation was considered significantly more when number information was highly salient in the broad context compared to contexts where it was less salient. This result was judged to be potentially compatible with the simple discourse account, whereas the strict identity account

predicts that such contextual manipulations should be ignored in the context of a linguistic antecedent and cannot readily account for these results.

Experiment 2 explored the possibility that the Experiment 1 results were actually compatible with the pure identity model because the same distributions of interpretations would be available in the corresponding contexts with the antecedent VP *in situ* in the target sentence. The experiment was a replication of Experiment 1 with the VPE sites for each condition replaced with a VP identical to the linguistic antecedent for that condition, if there was one. Comparison of the results for the two experiments indicated that the critical nonidentity effect observed in Experiment 1 was at least partially unique to ellipsis, meaning that the strict identity model’s difficulty in accounting for the Experiment 1 data persists.

The Experiment 1 results were judged to be potentially compatible with a simple discourse interpretive model for VPE. However, this relies on the assumption that there was a close mapping between the ratings collected for VPE in Experiment 1 and the general discourse availability of number-unmodified versus -modified propositions in each of the different discourse contexts. Experiment 3 aimed to test this assumption by collecting participants’ ratings of the coherence of the number-unmodified and -modified interpretation prompts in each discourse context, with the ratings taken as a proxy for the discourse availability of the two interpretations. Major qualitative differences in the results for Experiments 1 and 3 undermined the credibility of the simple discourse account, as the distribution of interpretations under VPE largely favored the interpretation supported by the linguistic antecedent in conditions where that interpretation was not judged to be more coherent in the discourse in Experiment 3.

Considered together, the results of Experiments 1 through 3 point toward a model of VPE interpretation in context that hybridizes an identity approach and a discourse approach. Experiment 1 showed significant differences in the interpretation of VPE based on changes in the information available in the comic strip context even in the presence of a viable linguistic antecedent, contra the predictions of the strict identity model of interpretation.

Experiment 2 verified that this was not due to ellipsis-independent facts of the interpretation of the corresponding sentences with *in situ* VPs. However, Experiment 3 indicated substantial differences between the interpretations available for VPE and the propositions that were salient in the discourse at large, whereas the pure discourse account predicts a one-to-one mapping between these.

The most illustrative experimental condition is the case of an unmodified linguistic antecedent and a salient context. The strict identity account predicts uniform interpretations for ellipsis sites with an unmodified antecedent in each of the unavailable, available, and salient contexts. However, the number-modified scalar interpretation received significantly better ratings in the salient context, exactly where number information was highly salient in the broader context. This suggests that the strict identity account must be augmented with a discourse-sensitive module to adequately account for the data.

The simple discourse account predicts that the polar and scalar interpretations of VPE should receive roughly equal ratings in this condition, since these interpretations are equally available in the discourse according to Experiment 3. Despite this prediction, it is actually the number-unmodified polar interpretation, which is exactly the interpretation predicted by the strict identity account, that is significantly preferred under ellipsis in this discourse context. This indicates that the simple discourse account needs an additional module that constrains interpretation based on the form of the linguistic antecedent beyond its contribution to discourse salience.

Thus, the experimental results constrain the hypothesis space for VPE interpretation on both the “purely identity” and “purely discourse” sides of the interpretive spectrum. Given this restriction, it seems likely that context-situated VPE interpretation is best modeled according to one of two more nuanced mechanisms identified in the previous chapter: a semantic identity/discourse account with complex interpretation heuristics, or a permissive syntactic identity account allowing for the construction of alternative antecedents.

The first plausible account holds that VPE interpretation is essentially discourse reference

– a search for an antecedent meaning in the discourse at large – but that the heuristics that guide interpretation are more complex than simply accessing propositions in proportion to their salience in the context. Hardt (1993) gives a partial sketch of what such a system would look like in exclusively linguistic discourses, with preferences for subject coreference and maintenance of particular clausal relations interacting with a general preference for the antecedent to be recent in the discourse.

The challenge for this account is to spell out the exact heuristics that guide interpretation in complex discourse contexts, where the linguistic antecedent and broader context interact to determine interpretation. In particular, the experimental results presented here indicate that the information from the linguistic antecedent must be considered beyond its mere contribution to discourse salience. Among the only accounts to explicitly address the antecedent-context interaction are Miller and Pullum (2013) and Miller and Hemforth (2014). They propose that VPE interpretation is discourse-driven, but that linguistic antecedents are much more available in processing than information from the broader context, giving rise to an apparent requirement for interpretation under linguistic identity.

This proposal seems to be on the right track in that it correctly privileges antecedent information over information from the broader context in VPE interpretation. However, it may conflict somewhat with the observation that there is not a straightforward mapping between the discourse availability of competing propositions and their availability under ellipsis, at least as measured by discourse coherence in Experiment 3. It may be the case that the coherence or salience measure reported in Experiment 3 is not an appropriate measure of the processing availability appealed to by Miller and colleagues. Alternately, this proposal may need to be strengthened to suggest a processing heuristic that explicitly privileges linguistic information beyond a simple measure of salience or discourse availability. Such a mechanism has precursors in the broader literature on anaphora; for instance, it is known that pronoun resolution involves balancing concerns of linguistic identity and resolution with respect to the broad context in ways that are not reducible to the salience of competing

potential referents (Kehler and Rohde, 2013).

It will likely take considerable additional theoretical and experimental work to fully spell out the heuristics that govern VPE interpretation in context under a discourse-driven model. One difficulty such an account makes is its relatively unconstrained predictions; in principle, it is possible to appeal to a large number of interacting heuristics, combined in the right proportion, and generate almost any interpretive pattern. The next chapter outlines a possible approach to testing this hypothesis in a principled way: probabilistic modeling of VPE interpretation in context. This approach allows for the predictions of complex interpretation mechanisms to be tested against empirical measures of interpretation, with model comparison techniques that penalize hypotheses that appeal to too many interpretation heuristics just to fit the data. While the models considered in the next chapter are just preliminary sketches of a context-situated VPE interpretation mechanism, they provide a skeleton for principled future investigation of discourse-driven VPE interpretation through the addition of additional modules to the interpretation mechanism.

The second plausible account is a permissive identity account allowing for the consideration of alternative antecedents beyond any local antecedent VPs. This contrasts with the strict identity account, which holds that interpretation must proceed according to the content of the antecedent, and must include an “escape hatch” for exophoric VPE, such as inference that the elided material is a deep anaphor like *do it* (e.g., Merchant, 2004). Under a more permissive account, exophoric VPE could be resolved by accommodating a covert antecedent based on the information that is salient in the nonlinguistic context, then interpreting the ellipsis site under linguistic identity with this antecedent. Likewise, effects of the broader context on interpretation in the presence of a linguistic antecedent would be attributed to the availability of additional alternative antecedents.

The construction of alternative VPE antecedents has several precursors in the literature, as accommodation or repair operations have been proposed several times to account for the acceptability of mismatch sentences that are ruled out under a syntactic identity account

for VPE (Fox, 1999, 2000; Arregui et al., 2006; van Craenenbroeck, 2012; Thoms, 2015). A major challenge for this model as an account of the current data is the lack of an overt trigger for accommodation in the non-exophoric conditions of the experimental paradigm. In the prior literature, a syntactic mismatch making the linguistic antecedent formally incompatible as an *in situ* interpretation for the ellipsis site is treated as the trigger for accommodation of an alternative antecedent. The accommodation is generally required to be “minimal” – that is, the overt antecedent should be altered just enough to make it syntactically compatible with the ellipsis site.

In the current experimental paradigm, accommodation could be argued to have a trigger in the exophoric conditions; since there is no linguistic antecedent whatsoever, there is a “mismatch” between the ellipsis site and the prior discourse, and an antecedent must be constructed from the information at hand. However, the unmodified- and modified-antecedent conditions do not have such a trigger. There is no obstacle to interpreting the VPE sites in these conditions with the antecedent VP *in situ*. Nevertheless, an effect of the broader context on interpretation was observed in the unmodified antecedent/salient context condition. An accommodation or repair account of this finding would represent a departure from the prior literature on consideration of alternative antecedents and deserves a theoretical explanation.

One possible trigger for accommodation in these contexts might be general monitoring on the part of listeners of the information that is apparently at issue or under discussion in the discourse at large. A recent approach to language comprehension, the noisy-channel model, holds that a major component of sentence understanding involves contending with noise and variability in the communicative signal (Levy, 2008; Gibson et al., 2013; Futrell and Levy, 2017; Ryskin et al., 2018). Work on the noisy-channel model investigates how listeners might contend with errors in both production, such as the speaker mis-speaking, and perception, due perhaps to simple mis-hearing or noise in the environment, to arrive at the signal the speaker intended to transmit. By hypothesis, listeners consider alternative

signals that are similar to the apparent percept, such as transformations involving deletion or insertion of a particular component or substitution of an alternative component. The listener's level of confidence that a particular signal was intended is conditioned by, among other things, the prior probability of that signal given the discourse and the probability of particular errors being made in production and perception.

(It is worth mentioning here that Bergen and Goodman (2015) have proposed a full noisy-channel account of VPE interpretation. Their proposal is that speakers, in the context of a viable linguistic antecedent, can minimize production costs to themselves by eliding antecedent-identical material under the assumption that the listener will be able to recover the meaning. The listener recognizes the sentence as syntactically incomplete and proceeds as though the missing material were present, but obscured by noise. The listener will generally arrive at the correct interpretation because the target sentence with the antecedent *in situ* is judged to be the most likely syntactically licit signal given the context.)

It seems plausible that this type of account could be extended to the current experimental paradigm. In the critical unmodified antecedent/salient context condition, the antecedent speaker's utterance would canonically be interpreted as indicating that number information is not at issue. However, there is salient information in the broader context suggesting that this speaker actually does care about the number of referents. Under the noisy-channel model, the listener would infer from this contextual support that the speaker has a high probability of transmitting a message where number information is at issue. When interpreting the apparent signal, the listener would also consider nearby alternative signals, potentially including an altered version of the actual antecedent that includes a numeral. Given the high probability of the signal addressing number information, and the relative plausibility of number information having been intended but excluded from the signal (by a speech error, by noise in the context, or for some other reason), the listener might conclude that they should treat a numeral-modified alternative as the intended antecedent, and subsequently interpret the ellipsis site under identity with this alternative antecedent rather than the one

that was intended in the discourse.

A related observation is that the critical nonidentity effect (the increased consideration of the scalar interpretation with an unmodified antecedent and salient context) came exactly in the condition where there was additional linguistic information available outside the antecedent VP, namely the visual representation of counting out the referents in the comic strip context. It is worth considering whether there is a possible augmented identity proposal that uses available linguistic material to construct alternative antecedents, but is actually not sensitive to information in the nonlinguistic context. Such a mechanism is plausible, although some numeric trends in the experimental data argue against it – for instance, the decreased consideration of the polar interpretation in the Experiment 1 unmodified conditions when changing from an unavailable context to an available one. In the next chapter, a similar non-significant numeric trend arises in a forced-choice paradigm, with slight modulation of interpretation in both the unmodified and modified antecedent conditions between the unavailable and available contexts. Future experimentation with additional attention to nonlinguistic context manipulations may verify that the interpretation mechanism is sensitive to information in the general context, including nonlinguistic information, and not just to linguistic information from outside the linguistic antecedent.

In summary, the results of Experiments 1 through 3 constrain the hypothesis space for VPE interpretation on two ends of an interpretive spectrum. A strict identity account is ruled out by contextual effects on interpretation with the same antecedent, but a simple discourse account is precluded by mismatches between the discourse availability of competing propositions and their availability under ellipsis. A discourse-driven account with a more articulated set of interpretation heuristics and a more permissive identity account allowing construction of alternative antecedents remain viable hypotheses. Future work may help refine the predictions of these hypotheses and adjudicate between them. The next chapter presents additional consideration of the interpretation for VPE mechanism using a modeling methodology that may become a useful tool for testing the more nuanced predictions of each

hypothesis.

3.5.2 Comparing VPE and *do that*

Experiment 4 investigated the interpretation of the verbal anaphor *do that* using the experimental paradigm from Experiment 1. This served partially as a check of the experimental paradigm, as *do that* canonically would be expected to be more sensitive to information in the discourse context than VPE (Hankamer and Sag, 1976; Tanenhaus and Carlson, 1990; Mauner et al., 1995; but cf. Murphy, 1985), but also as an exploration of how to systematically compare the interpretive mechanisms for multiple constructions and draw deeper conclusions for the theory of anaphora and context dependence.

Considering the results of Experiment 1 (VPE) and Experiment 4 (*do that*) together, there were some striking qualitative similarities in the results for all three antecedent types. With no linguistic antecedent, both constructions unsurprisingly showed a preference for the polar interpretation with an unavailable context, where number information was not available in the nonlinguistic context. In both experiments, the ratings for the polar interpretation decreased and the ratings for the scalar interpretation increased as a function of the salience of number information in the comic strip context as it changed from unavailable to available to salient.

With an unmodified antecedent, the polar interpretation was preferred for both expressions when number information was not highly salient in the nonlinguistic context (i.e., unavailable and available). The results for both experiments showed that the ratings of the polar interpretation did not change significantly in response to the comic strip context manipulation. However, for both constructions, the ratings for the scalar interpretation were significantly higher in the salient context, where number information is highly available, than in the unavailable or available contexts, where it is less available.

Finally, with a number-modified linguistic antecedent, both expressions showed a global preference for the scalar interpretation. This is again unsurprising; since the speaker of the

linguistic antecedent specifically mentioned number information in their utterance, it would be unusual for listeners to conclude that the interpretation of the reply should not itself be number-modified.

Despite the overall similarities in the trends, however, there are some telling differences in the results for the two experiments that suggest that *do that* is more sensitive to information from the broader discourse context than VPE – in other words, that *do that* is a “deeper” anaphor than VPE.

In the exophoric conditions, both constructions show improvement for the scalar interpretation and degradation for the polar interpretation as number information becomes more salient in the comic strip context. For VPE, the polar interpretation was preferred in the unavailable context, the two interpretations were rated approximately equally in the available context, and the scalar interpretation was preferred in the salient context. For *do that*, the polar interpretation similarly was preferred in the unavailable context. However, in the available context, the change in ratings for the two interpretations was already substantial enough that the scalar interpretation was preferred, a result which carried over to the salient context.

The different results for the two expressions in the available context points to a difference in sensitivity to information from the broader discourse context. With number information only somewhat available in the comic strip context, interpreters of VPE were indifferent between the polar and scalar interpretations, whereas interpreters of *do that* significantly preferred the scalar interpretation. For VPE, only when number information was highly salient, in the salient comic, was the scalar interpretation preferred. This suggests that *do that* is more sensitive to contextual information than VPE, since a contextual effect that was only observed for highly salient information in VPE interpretation was observed when the same information was less salient in the *do that* experiment.

In addition, the mismatches in interpretation between exophoric VPE and exophoric *do that* are suggestive evidence that exophoric VPE interpretation does not involve inference

that the elided material is *do that*. Recall that Merchant (2004) proposes that exophoric VPE interpretation requires an exceptional mechanism by which listeners infer that the elided content is a deep anaphor that can be freely interpreted with respect to the discourse. Since the interpretation preferences for exophoric VPE and exophoric *do that* were not isomorphic, it seems that this model may not be adequate with *do that* as the elided anaphor. However, it remains possible that hearers of exophoric VPE infer that the elided material is a different anaphor with different interpretation preferences; in fact, Merchant specifically mentions *do it*, not *do that*, as the elided deep anaphor. Thus, further research is warranted to determine if the elided deep anaphor analysis of exophoric VPE is tenable with an anaphor other than *do that*.

With an unmodified antecedent, both VPE and *do that* showed approximately equal ratings for the polar interpretation with all three comic strip contexts, whereas the scalar interpretation received higher ratings in the salient context than in the unavailable context. Interestingly, however, VPE showed an across-the-board preference for the polar interpretation in all three contexts, while the scalar interpretation was marginally preferred in the salient context for *do that*. Again, this finding suggests that similar increases in salience for information (in this case, number information) in the broader discourse have an outsized impact on the interpretation of *do that* compared to the interpretation of VPE. These findings for the exophoric and unmodified conditions appear to confirm the widely cited intuition that *do that* is a “deeper” anaphor than VPE and as a result is more sensitive to changes in the information that is salient in the broader discourse context beyond the linguistic antecedent. This suggests that the experimental paradigm appropriately detects the interactions between linguistic antecedents and broader contexts in the interpretations of context sensitive verbal expressions.

It is also worth comparing the results for *do that* to the Experiment 3 coherence results to explore what type of interaction with the discourse context at large is typical for a deep anaphor. Again, there are substantial similarities in the results for the two experiments.

The Experiment 3 results show degradation of the polar interpretation as a function of the salience of number information in the comic strip context, like the *do that* results. The coherence results also showed the characteristic boost for the scalar interpretation in the unmodified/salient condition compared to the unavailable and available contexts, although this jump did not attain significance in that experiment.

However, there are some qualitative differences in the results for the two experiments that suggest that the interpretation of *do that* may not entail simple recovery of the most salient propositions in the discourse context (simple discourse model). In the modified-antecedent conditions, *do that* showed a preference for the scalar interpretation in all three context conditions, whereas the coherence results showed indifference between numeral-modified and -unmodified propositions in all three contexts. This suggests that *do that* may be somewhat more sensitive to the content of the linguistic antecedent than predicted by a simple discourse account of *do that* interpretation. This is because the antecedent contains a numeral in these conditions, which seems to have promoted consideration of the scalar interpretation in the *do that* experiment beyond what the antecedent contributes to the discourse at large.

In the unmodified-antecedent conditions, the coherence results showed indifference between the two interpretations in a salient context, whereas *do that* showed a marginal preference for the scalar interpretation in the same condition. It is difficult to know how much weight to ascribe to this finding for *do that* since it did not attain significance, but it again suggests a divergence between *do that* and the general discourse availability of the two interpretations. Here, however, the unexpected preference (from the perspective of the simple discourse account) is in favor of information from the broader discourse context rather than information from the linguistic antecedent. This is difficult to reconcile with the modified-antecedent results, which suggested that *do that* was more sensitive to linguistic antecedent information than predicted.

Finally, in the exophoric conditions, it is notable that there was a significant preference for the scalar interpretation in the available context, whereas the coherence results showed

indifference between the two interpretations in this condition. This finding is particularly difficult to explain, as in the absence of a linguistic antecedent, interpretation must proceed entirely along the lines of the information available in the comic strip context. Thus, it is not clear why there should be a difference between the general discourse availability of the two interpretations and the interpretations that are available under *do that*.⁵

The experimental findings for VPE and *do that* interpretation are potentially interesting from the perspective of the broader theory of anaphora. Conventionally, VPE and *do that* have been treated as lying on opposite sides of the deep-surface divide. VPE is described as a surface anaphor, meaning it is interpreted under linguistic identity with an antecedent. By contrast, *do that* is a deep anaphor, freely receiving its interpretation from the discourse context at large.

The results of the present experiments suggest that the surface-deep divide may not be as stark as previously hypothesized. VPE was shown to be sensitive to information in the broader context, even when used with a viable VP antecedent. Likewise, interpretation of *do that* was not simply a matter of accessing the most salient proposition in the context; for instance, in the modified antecedent conditions, an antecedent-identical interpretation was more available for *do that* than its relative coherence with the prior context would suggest.

The observation that neither construction is interpreted fully along identity lines or fully along discourse lines suggests that they may sit on an interpretive continuum rather than on two sides of a categorical divide. Given that a discourse-driven interpretive mechanism with complex heuristics was highlighted as a possible model of VPE interpretation, it is possible that both constructions are actually “deep” in the sense that their interpretations holistically consider the discourse at large. Different constructions may introduce different form-specific

5. While these experiments have not been discussed yet, Experiments 5 and 6 from the next chapter suggest that a number-modified proposition is actually significantly more available in the unmodified/salient context. If this is correct, then the marginal preference for the scalar interpretation for *do that* actually underestimates the strength of the scalar interpretation compared to the predictions of the simple discourse account. This may suggest that *do that*, like VPE, exhibits interpretive patterns whereby concerns of general discourse salience are constrained by the form of the linguistic antecedent, although the magnitude of this effect is, unsurprisingly, less for *do that* than for VPE.

biases for accessing linguistic versus broader context information in interpretation, leading to the impression that they are relatively more “surface” or “deep.”

On the other hand, the possibility remains that the grammar of VPE interpretation actually does proceed under linguistic identity, with a separate mechanism like accommodation or repair accounting for the effect of context. In this case, there might actually be a grammatical divide between surface anaphors like VPE on the one hand and deep anaphors like *do that* on the other. The accommodation/repair operation for VPE makes it appear impressionistically “deeper” than the grammar suggests, while the heuristics for *do that* interpretation make it appear more “surface” than the most salience-sensitive extreme. The empirical consequences of this proposal are largely similar to the previous ones, but the underlying theoretical architecture and the implications for the theory of anaphora are quite different.

The results presented here do not clearly indicate whether these two constructions fall on an interpretive cline or on different sides of a categorical divide in terms of their interpretive mechanism, but the results suggest that the empirical situation is, unsurprisingly, more complicated than the neat picture that is sometimes depicted in the literature. Further research on individual constructions may yield important conclusions – for instance, whether or not VPE interpretation involves a syntactic module. However, it seems likely that systematic investigation of a large number of anaphoric constructions will be necessary to draw deeper conclusions about the nature of context dependence. This chapter provides a blueprint for a research program on this path by applying an identical experimental paradigm to two constructions; this or a similar paradigm could also be extended to other anaphors such as *do it* or *do so*. The modeling methodology presented in the next chapter may also extend the methodological toolkit for testing and comparing the predictions of different models for the interpretation of various verbal anaphors.

CHAPTER 4

A PROBABILISTIC MODEL OF VPE INTERPRETATION IN CONTEXT

The previous chapter presented experimental results that were argued to constrain the hypothesis space for the mechanism of context-situated VPE interpretation. A critical effect of the broader discourse context on interpretation in the presence of a constant antecedent ruled out a “strict” syntactic identity account that ignores additional contextual information when an antecedent is available. At the same time, there was an apparent mismatch between the availability of propositions in the discourse at large and the degree to which they were considered as ellipsis interpretations, indicating that a “simple” discourse model under which interpretations are considered in proportion to their salience in the discourse is also inadequate.

Two more complex models of interpretation were determined to be plausible: a discourse-driven account featuring complex interpretation heuristics, and a permissive identity account allowing for the consideration of alternative antecedents. Impressionistically, these accounts are ruled in because they can “hybridize” the predictions of the two more extreme accounts. The complex discourse account can shift the interpretation distribution predicted by the simple discourse account toward the predictions of an identity account if the interpretation heuristics favor the use of antecedent information beyond its contribution to general salience. Likewise, the permissive identity account can shift the predictions of a strict identity model toward those of a discourse-driven account by allowing interpretation with respect to alternative antecedents that are supported by the broader context.

This chapter aims to construct further evidence for this conclusion, as well as to sketch a new tool that can be used to investigate the interpretive mechanisms of context-dependent constructions like VPE. The main contribution is the construction and comparison of several different probabilistic models of context-situated VPE interpretation. These models

transform probabilities of accessing a particular interpretation in the prior discourse into predicted probabilities of accessing a particular interpretation for an ellipsis site, which in turn can be compared to empirical interpretation probabilities from an experimental study.

Three main classes of model are considered: “identity” models schematizing interpretation only with respect to a linguistic antecedent, “discourse” models schematizing interpretation with respect to discourse salience alone, and “hybrid” models combining these two strategies in proportion. Model fitting and comparison reliably indicates that the hybrid models provide the best fit for the empirical facts of VPE interpretation in context, even taking into account their complexity relative to the identity and discourse models (i.e., overfitting penalties).

In addition to this main modeling activity, the chapter provides experimental replication of the findings of the previous chapter for VPE interpretation and its connection to the discourse availability of competing interpretations. Since the previous chapter used a Likert scale rating task, there was not a clear mapping between participants’ responses in the experiments and their probability of accessing particular propositions in the discourse context or as ellipsis interpretations. Thus, the chapter opens with Experiments 5 through 7, which assess the empirical availability of number-unmodified and number-modified propositions in the prior discourse context and as VPE interpretations in the experimental paradigm from the previous chapter. The Likert rating task is changed to a forced choice between multiple interpretations. This allows the results to be expressed as proportions, allowing for their use as empirical prior and posterior probabilities in the modeling section that follows.

4.1 Experiments 5 and 6: Empirical priors – Availability of competing meanings before VPE

The focus of Experiments 5 and 6 is to empirically estimate the prior availability of the previous chapter’s polar and scalar meanings in each experimental condition before the

use of the VPE reply. To do this, the experiments elicit participants' assessment of the likelihood of different possible communicative goals in Experiment 1-style discourses with the VPE reply removed. Some of these goals are taken as “priors” for the polar and scalar interpretation – that is, as indicating that number information is or is not at issue in the discourse at large. In order to estimate the probability that participants think a particular interpretation is at issue, the experiments use a forced-choice methodology. This allows the results to be expressed as proportions of trials in which each meaning was selected, enabling estimation of a probability distribution.

The experiments are intuitively similar to Experiment 3 in that they seek to explore the connection between the interpretation of the VPE reply and the information available in the prior discourse at large. However, these experiments are intended to more directly assess the information that is available in the discourse prior to the VPE utterance, and participants did not see the VPE utterance at all in this experiment. In essence, the goal is to determine the probability in each condition that participants will think that number information is at issue before they even know that the next utterance will use a VPE construction, let alone that it might or might not crucially turn on the inclusion of number information.

4.1.1 Design, methods, and materials

The design and materials of the experiments were largely similar to Experiment 1. The experiment retained a 3x3 design with the same comic strip context (unavailable, available, salient) and antecedent (exophoric, unmodified, modified) manipulations as Experiment 1, and used the same scenarios, utterances, and comic strip contexts, minus the VPE reply.

Three major alterations were made. First, the VPE reply was deleted, so each trial consisted of a scenario title, the comic strip context, and an antecedent utterance in the conditions where appropriate. Second, instead of being prompted with a single interpretive choice and providing a Likert rating of it, participants were presented with all possible response choices and required to make a forced choice between them.

The third change was to alter the interpretation task and the choices available. In these experiments, participants were asked to assess what they thought was the most likely “mental state” of the speaker of the antecedent sentence in each scenario. Three mental states were considered. In the **zero** mental state, the speaker did not want to interact with any of the referents associated with the scenario (*The son does not want to buy any candy bars.*). In the **indifferent** mental state, the speaker wanted to do something with the referents, but was indifferent to the number of referents (*The son wants to buy candy bars, but doesn’t care how many.*). In the **specific** mental state, the speaker had a particular number of referents they intended to interact with (*The son wants to buy a specific number of candy bars.*).

Below, the proportion of indifferent responses will be treated as a prior for the polar VPE interpretation, and the proportion of specific responses will be treated as a prior for the scalar VPE interpretation. The indifferent response explicitly encodes that number information is not at issue in the antecedent speaker’s conceptualization of the discourse. If VPE interpretation proceeds entirely according to the availability of different meanings in the discourse, then a reply like *We can’t* under such conditions should target only the possibility of taking some course of action, without respect to number information. Similarly, the specific response explicitly encodes that number information is at issue in that participant’s mental representation of the discourse, and if VPE interpretation is driven exclusively by discourse salience, then *We can’t* is likely to address number information, as well.

The zero interpretation is not considered to be a prior for any viable interpretation of the VPE reply. That is, if the son in the candy bar scenario intends for his utterance (or lack thereof) to encode that he does not want any candy bars, it is not clear what, if anything, a reply of *We can’t* would mean. However, it seemed clear during experimental design that this mental state might be preferred in some of the experimental conditions. In particular, in the exophoric/unavailable condition, where number information is not salient in the comic and there is no antecedent, there is not strong evidence that the character wants to interact with the referents, and the zero mental state might be most accessible for some participants.

The only difference between Experiments 5 and 6 is that Experiment 5 included the zero mental state as a response choice and Experiment 6 excluded it, with the latter providing only the indifferent and specific options. Experiment 5 is likely to be the more natural experiment for participants, as it includes mental states that should nearly exhaust the range of possible at-issue content. Experiment 6 might be less natural for participants if they would otherwise prefer the zero mental state in a particular condition, but cannot select it. However, Experiment 6 has the advantage of a more straightforward mapping onto the VPE interpretation task, as there are only two interpretation options that are taken to correspond directly to the polar and scalar interpretations from the VPE tasks.

The interpretation prompt was shown below the antecedent utterance when it was present, and below the comic strip context otherwise. Participants were asked, *Based on the scenario above, which of the following do you think is most likely?*, followed by the two or three possible mental states in a random order by trial. To provide their response, participants could either click on their preferred choice or press the corresponding number key on their keyboard. Figure 4.1 shows how a representative trial in Experiment 5 appeared in the Ibex Farm experiment. A sample screen from Experiment 6 would be identical, except that the zero mental state (*The son does not want to buy any candy bars*) would not be an option.

The practice trials and fillers from Experiment 1 were adjusted to have the same format as the critical trials, with the VPE reply removed and the prompt altered to a forced choice between different mental states attributed to a character. The procedures for informed consent, recruitment, and payment were identical to Experiment 1.

4.1.2 Participants

227 participants took part in Experiment 5. Of these, 103 were female and the mean age was 34.8 years. The data from 2 participants was excluded from analysis because they did not clearly identify themselves as native speakers of English in the demographic survey. 14 additional subjects had mean reaction times per trial of under 2000 ms, indicating they were

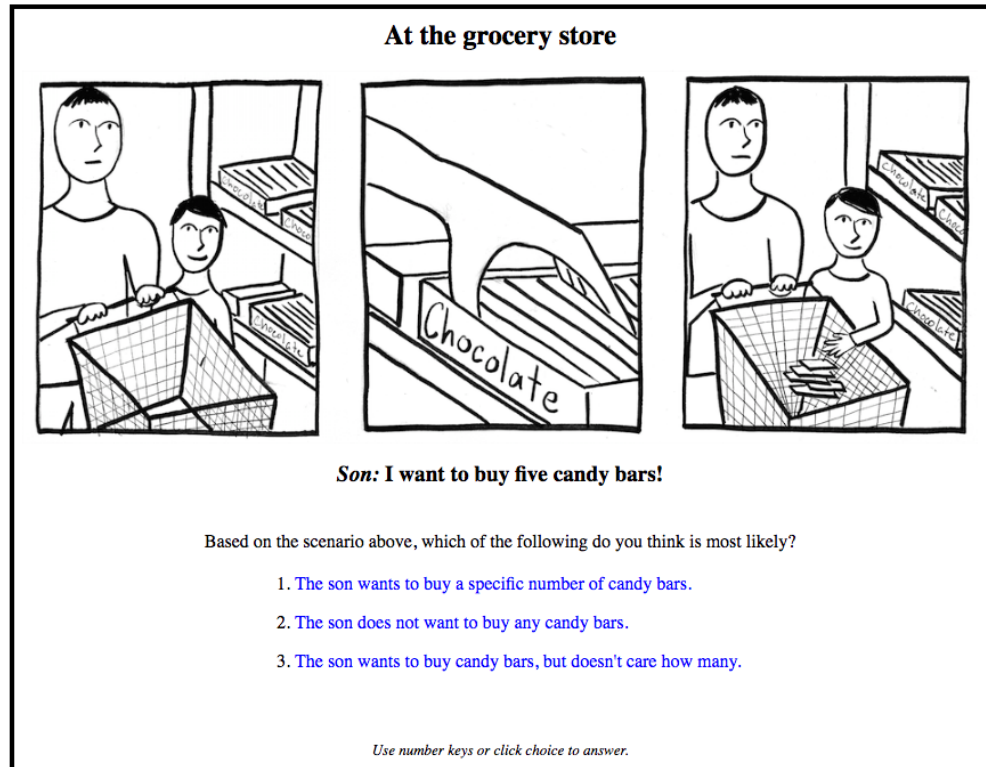


Figure 4.1: Sample screen from Experiment 5. A screen from Experiment 6 would be identical, except that the zero mental state (*The son does not want to buy any candy bars.*) would not appear.

not fully attending to the experiment, and their data was excluded. Finally, the data from an additional 10 participants was excluded because they incorrectly answered at least two out of three fillers that were designed to test for attention. The mean completion time for participants whose data was analyzed was 7 minutes, 12 seconds, and participants received USD 1.00 for completing the experiment.

224 participants took part in Experiment 6; 111 were female and the mean age was 35.6. Data were excluded from analysis for the same reasons as Experiment 5; 6 participants were excluded for their native language response, an additional 4 for their mean response time, and a further 6 for their responses to filler trials. The mean completion time was 6 minutes, 18 seconds, and participants received USD 1.00 for participating.

4.1.3 Results

Figure 4.2 shows the results of Experiment 5 with the responses expressed as the proportion of trials in each condition that each of the three mental states (zero, indifferent, specific) were selected. Because the primary goal of these experiments is to assess the prior availability of the polar and scalar VPE interpretations in each context, estimated using a proportion, Figure 4.3 also shows the proportions of indifferent and specific responses in each condition of Experiment 5 after discarding all zero responses. Finally, Figure 4.4 shows the results of Experiment 6 expressed as proportions of indifferent and specific responses by condition.

The results appear largely unsurprising when given a visual inspection. First, the zero mental state in Experiment 5 received some consideration in the conditions with unavailable comic strips, and especially the unavailable/exophoric condition. This condition was identified above as the scenario where the zero mental state would be most natural, so this result is not troubling. In the other 8 conditions, it was selected on fewer than 11% of trials and is not likely to have impacted assessments of the relative strength of the indifferent and specific mental states with respect to one another.

Turning to the indifferent and specific mental states, the results consistently point to a large effect of switching from an available to a salient comic, with the specific mental state appearing to be highly preferred in all conditions with a salient comic. This effect is familiar from the VPE interpretation experiments, where the inclusion of the linguistic numeral in the salient comic consistently had a larger effect than the change in nonlinguistic information between the unavailable and available comics.

With no antecedent or an unmodified antecedent, the indifferent mental state was numerically preferred in the context of an unavailable or available comic, while the specific mental state was numerically preferred with a salient comic. With a modified antecedent, the specific interpretation was numerically preferred across the board. This is unsurprising, since the linguistic antecedent includes a numeral in these conditions and number information should be treated as highly at issue. This effect was reinforced with a salient comic,

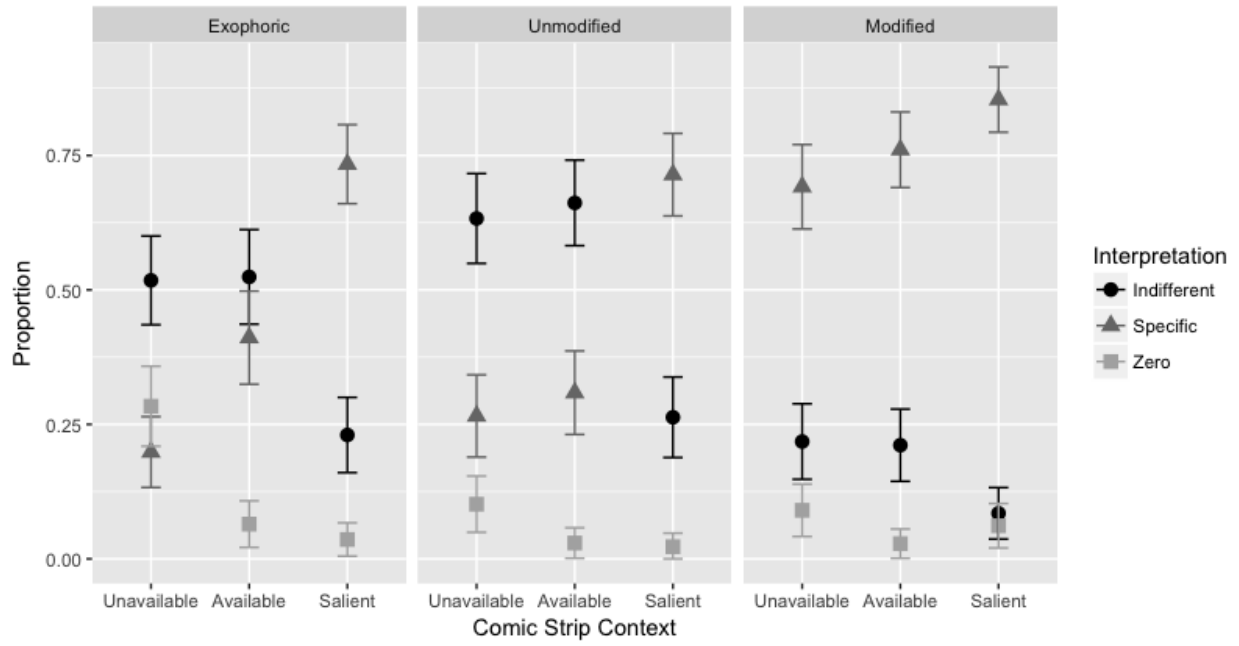


Figure 4.2: Experiment 5 results. Frame split: Antecedent. Error bars: 95% confidence interval.

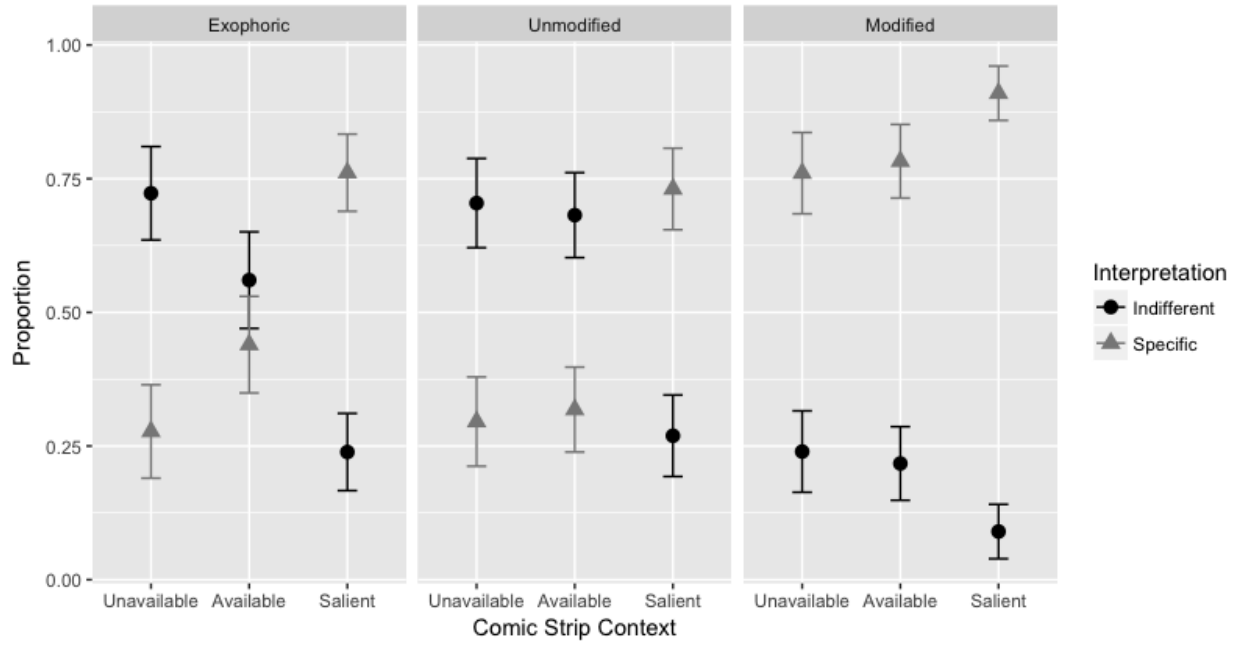


Figure 4.3: Experiment 5 results with zero interpretation removed. Frame split: Antecedent. Error bars: 95% confidence interval.

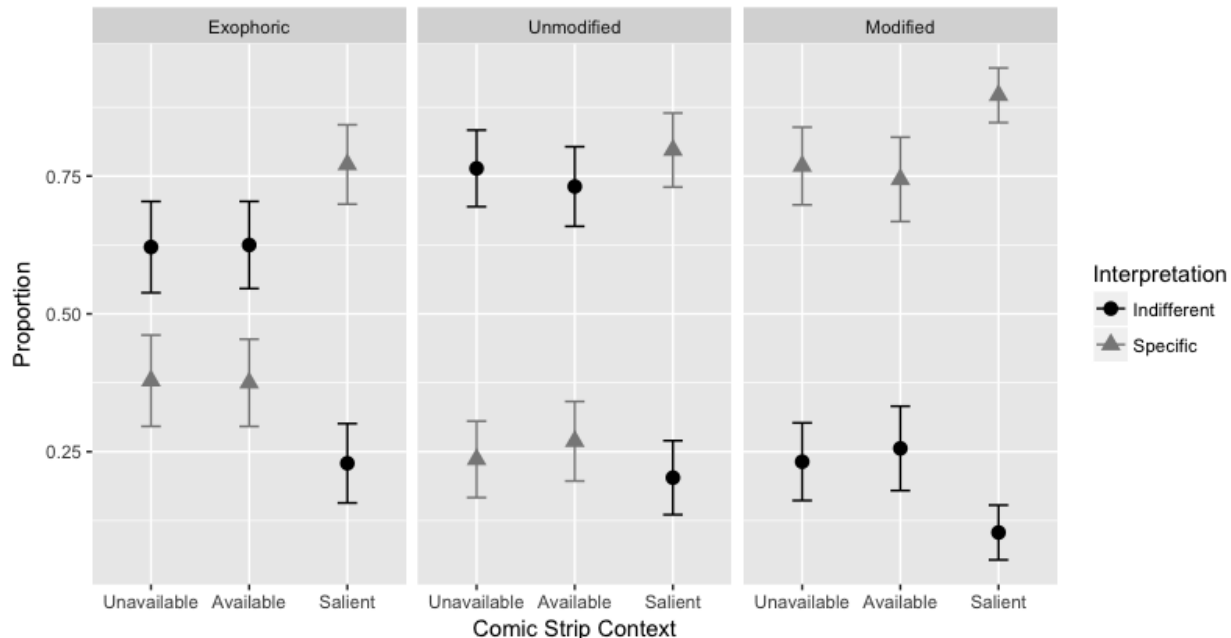


Figure 4.4: Experiment 6 results. Frame split: Antecedent. Error bars: 95% confidence interval.

where the strength of the specific interpretation increased even more over the other modified-antecedent conditions.

4.1.4 Analysis

The main goal of these experiments was to obtain numeric estimates of the probability that listeners find number information to be at issue or not at issue in each of the discourse contexts from Experiment 1 before the use of the VPE reply. As such, the raw proportions themselves are of interest, and the results will not be further analyzed at much length.

However, one straightforward and potentially useful analysis is to test whether, in each experimental condition, either the indifferent or specific mental state was significantly preferred over the other. Comparing the preference by condition in these experiments to the preference between the polar and scalar VPE interpretations by condition can provide some qualitative insight into the ability of different interpretive strategies to account for the mapping between prior discourse availability and VPE interpretation.

To test this, the Experiment 5 results with the trials with zero responses discarded and the Experiment 6 results were analyzed separately, with indifferent responses coded as 1 and specific responses coded as 0. For each set of results, a logistic mixed model with random intercepts for item was constructed for the data from each context/antecedent combination.¹ Note that no random effect was included for subject because each subject saw each condition combination at most one time.) This tests whether, separate from any effects unique to particular experimental items, the proportion of indifferent responses was significantly different from 0.5 in each condition. If the proportion is significantly different from 0.5, it indicates that either the indifferent mental state or the specific mental state received significantly more responses in that condition.

For the Experiment 5 results with zero responses removed, the proportion of indifferent responses was significantly different from 0.5 in the following conditions for the accompanying significance levels: unmodified/unavailable, $p < .05$; modified/unavailable, $p < .01$; exophoric/salient, unmodified/salient, modified/available, modified/salient, $p < .001$. The difference from 0.5 was marginal ($p < .1$) in the unmodified/available condition and did not attain significance in the exophoric/unavailable ($p > .1$) or exophoric/available ($p > .3$) conditions. (The non-significant result in the exophoric/unavailable condition is striking given the appearance of Figure 4.3, but may be driven by the relatively large number of zero responses that were removed, leaving a smaller number of observations relative to the other conditions.)

For the Experiment 6 results, the proportions were significantly different in the following conditions at the accompanying significance levels: unmodified/unavailable, unmodified/available, $p < .01$; exophoric/salient, unmodified/salient, modified/unavailable, modified/available, modified/salient, $p < .001$. The difference was not significant in the exophoric/unavailable or exophoric/available conditions (p 's $> .1$).

1. **glmer** model specification: $\text{Response} \sim 1 + (1 \mid \text{Item})$.

4.1.5 Discussion

The primary purpose of Experiments 5 and 6 was to collect empirical estimates of the proportion of times participants access number-unmodified versus number-modified propositions in the Experiment 1 paradigm prior to the VPE reply utterance. These proportions will be used as empirical prior probabilities in the modeling section below. The next section presents an experiment gathering empirical posterior probabilities, i.e., VPE interpretations.

Beyond this mechanistic goal, however, the experiments also provided a clearer picture of the availability of competing information in the discourse context than Experiment 3 did. The results of the two experiments agree that a number-unmodified proposition is significantly more available with an unmodified antecedent and an unavailable or available context. Likewise, they agree that a number-modified proposition is more available with either a salient context, regardless of antecedent, or a modified antecedent, regardless of context.

These findings validate the conclusion from the previous chapter that a simple discourse model cannot account for the facts of VPE interpretation in context. The most important condition to this conclusion remains the unmodified/salient condition, where a number-modified proposition is significantly more available in the context at large, but the antecedent-identical polar interpretation, which does not include number information in the ellipsis interpretation, is significantly more available under ellipsis.

4.2 Experiment 7: Empirical posteriors – Forced-choice VPE interpretation

Experiment 7 is a replication of Experiment 1 with one major alteration. In Experiment 1, after viewing the scenario and reading the VPE reply, participants were prompted with only one of the interpretations, polar or scalar, and provided a rating of the likelihood of the speaker intending that interpretation on a Likert scale. In the present experiment,

participants were presented with the two alternatives simultaneously, and made a forced choice about which interpretation was more likely to be the intended one. The alteration was made so the results could be expressed as a proportion of trials on which the polar versus scalar interpretation was selected in each condition; these proportions are in turn used as estimates of the probability of accessing a particular interpretation in the modeling section presented below.

4.2.1 Design, methods, and procedure

The design of the experiment was generally identical to Experiment 1, with a 3x3 cross of comic strip context (unavailable, available, salient) and antecedent (exophoric, unmodified, modified). As noted above, the exception is that the current experiment eliminates the polar/scalar manipulation, presenting the two interpretations simultaneously.

As before, on a trial screen, participants saw a scenario title, the comic strip context, the antecedent utterance when appropriate, and the VPE reply on the screen simultaneously. The polar and scalar interpretations were presented below a prompt asking participants, for example, “Based on the scenario above, do you think it is more likely that **the father** meant:”. Participants could reply by clicking on their preferred interpretation, or by pressing the number key corresponding to their choice on their keyboard. The order in which the two interpretations were presented was randomized by trial. Figure 4.5 shows the layout of a representative trial.

The procedures for informed consent, recruitment, and payment were identical to Experiment 1. The practice trials and fillers were altered to a forced choice task to correspond to the modification in the critical task.

4.2.2 Participants

198 participants (85 female, mean age 34.4) participated in the experiment. 13 subjects were excluded from analysis because they failed to correctly answer at least two out of three

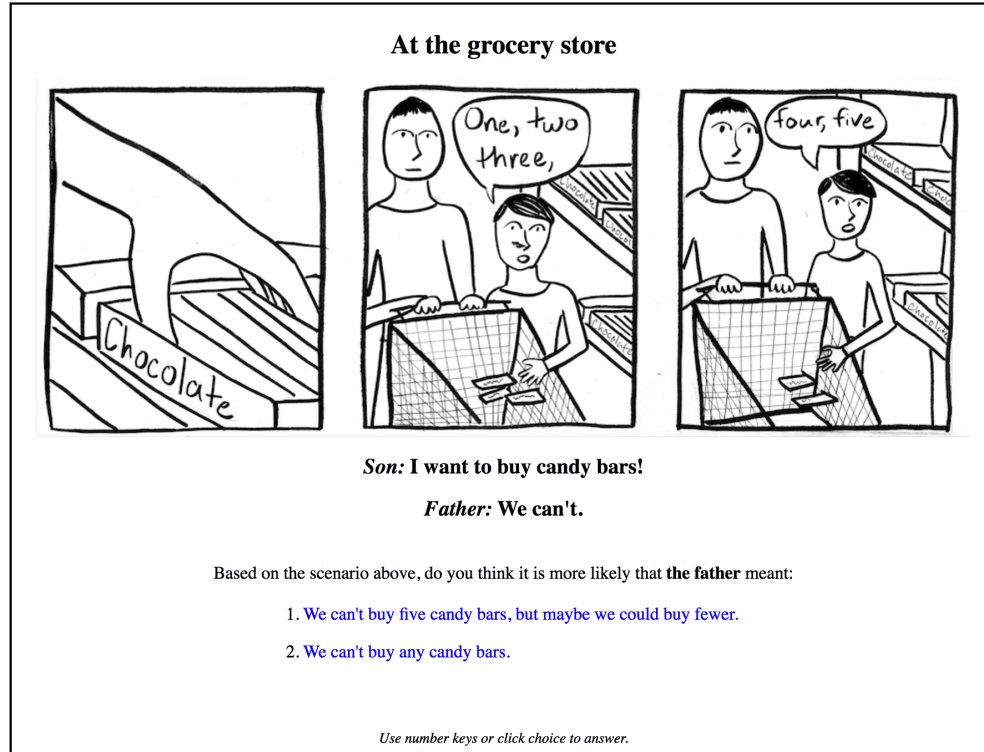


Figure 4.5: Sample screen from Experiment 7.

filler items that were designed to measure participant attention. 4 additional subjects were excluded from analysis because they failed to clearly identify themselves as native English speakers in the demographic survey. The mean completion time for participants whose data was analyzed was 7 minutes, 21 seconds. Participants were compensated with USD 1.00 for participating in the experiment.

4.2.3 Results

The proportions of polar and scalar responses by condition are shown in Figure 4.6.

A visual examination of the plot suggests that both the antecedent manipulation and the comic strip context manipulation had an effect on the proportion of polar versus scalar responses. In all three antecedent conditions, the proportion of times the polar interpretation was selected appears to decrease as a function of the number information available in the

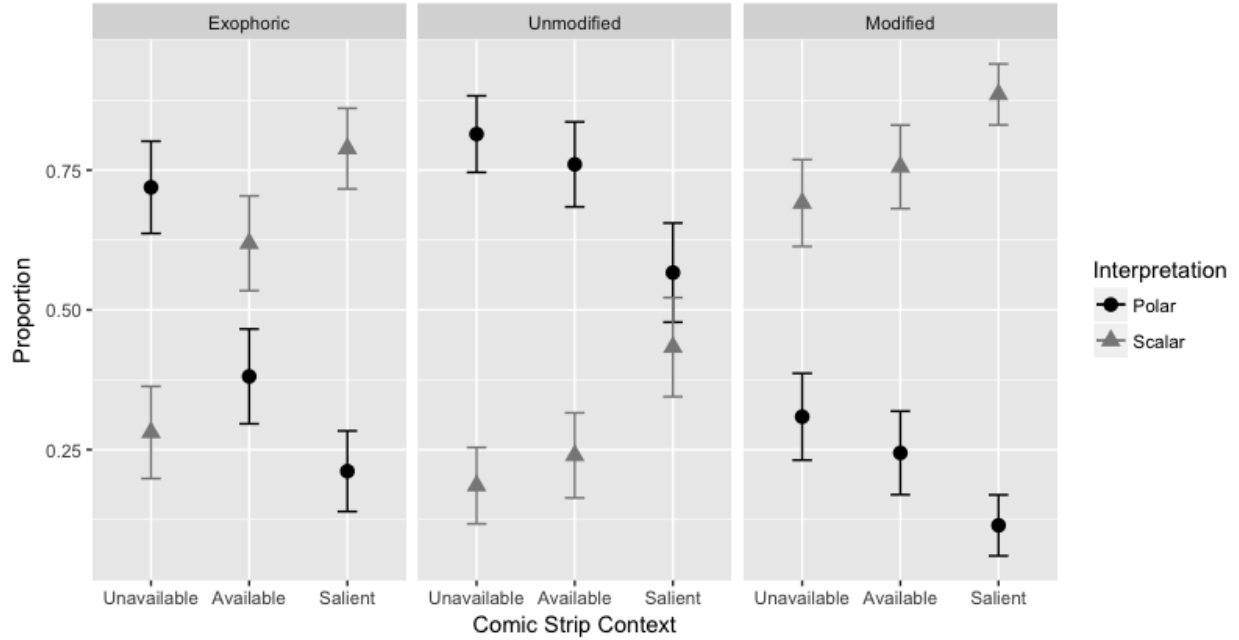


Figure 4.6: Experiment 7 results. Frame split: Antecedent. Error bars: 95% confidence interval.

comic strip context. However, this effect is mediated by the antecedent. In the exophoric conditions, a preference for the polar interpretation in the unavailable context gives way to a preference for the scalar interpretation in the available and salient contexts. In the unmodified conditions, the polar interpretation is always preferred, although its advantage erodes with increasing number information in the comic strip context, and the proportions of polar and scalar responses are close with a salient context. With a modified antecedent, the scalar interpretation is preferred even with an unavailable context. In an interesting departure from the Experiment 1 results, the strength of the scalar interpretation with a modified antecedent appears to increase as a function of context, with the preference for the scalar interpretation becoming even stronger in the available and especially the salient contexts.

4.2.4 Analysis

The response data were coded with polar responses as 1 and scalar responses as 0. Two logistic mixed effects regression models were constructed. The first model predicted the response data with fixed effects for comic strip context, antecedent, and their interaction. This model featured the maximally convergent random effects structure: random intercepts for participants, and a random intercept for item and random slopes for item in context and antecedent.² The second model had the same specification with the exception of the interaction of comic strip context and antecedent. Model comparison showed that the model with the interaction was a significantly better fit for the data ($p < .01$). On the basis of the significant interaction, further analysis of the effect of comic strip context on the proportion of polar responses was conducted on a by-antecedent basis.

For each antecedent condition, two types of analyses were performed. First, logistic mixed-effects regression models for the data from each antecedent condition were constructed with a main effect of context, random intercepts for participant, and random intercepts for item and random slopes for item in context.³ This analysis examined whether there was a gradual change in the polar response proportions when contextual support for number information increased from unavailable to available and then to salient. For this analysis, contextual support was coded using backward difference coding, with the unavailable context coded as the baseline level. With this coding scheme, the proportion of the polar response at each context level was compared to the proportion at the previous level.

In the exophoric conditions, the proportion of polar responses in the available condition was significantly different from the proportion in the unavailable condition, and the proportion in the salient condition was significantly different from that in the available condition (p 's $< .01$). In the unmodified antecedent conditions, the proportion of polar responses was

2. **glmer** model specification: $\text{Response} \sim \text{Context} * \text{Antecedent} + (1 \mid \text{Participant}) + (1 + \text{Context} + \text{Antecedent} \mid \text{Item})$. To attain model convergence, the random participant effects were selected for greater effect reduction due to the sparse nature of by-subject responses.

3. **glmer** model specification: $\text{Response} \sim \text{Context} + (1 \mid \text{Participant}) + (1 + \text{Context} \mid \text{Item})$.

not significantly different in the unavailable and available conditions ($p > .2$), but the proportions in the salient and available conditions were significantly different ($p < .001$). Similarly, in the modified conditions, the proportions were not significantly different for unavailable and available strips ($p > .3$), but were significantly different between the available and salient comic strip contexts ($p < .05$).

The second analysis was to determine whether in each condition, participants significantly preferred one interpretation over the other. A logistic mixed model with random intercepts for item was constructed for the data from each context/antecedent condition.⁴ The proportion of polar responses was significantly different from 0.5, indicating a significant preference for one of the two interpretations, for every condition besides unavailable/salient ($p > .1$) at the following significance levels: exophoric/available, $p < .05$; exophoric/unavailable, $p < .01$; all other conditions, $p < .001$.

4.2.5 Discussion

The results of Experiment 7 largely replicated the results of Experiment 1, albeit with some interesting differences. In the exophoric conditions, interpretation was largely a function of the information available in the comic strip context. The proportion of times the polar interpretation was selected increased as a function of the comic strip context manipulation, while the proportion of scalar responses likewise decreased. The polar interpretation was reliably preferred with an unavailable context, but the scalar interpretation was reliably preferred with an available and salient context. The finding for the available context is notable, since in Experiment 1 the Likert ratings for the two interpretations in this experiment were nearly identical. In addition, it is interesting that the polar interpretation was reliably preferred with an unavailable context when a number-modified proposition was not determined to be significantly more available in the context in Experiments 5 and 6. This may indicate some

4. **glmer** model specification: $\text{Response} \sim 1 + (1 \mid \text{Item})$. Note that no random effect was included for subject because each subject saw each condition combination at most one time.

unreliability in the mapping between the prompts of the two experimental paradigms. However, this condition is discursively odd in each experiment, as it assesses what a character wants to do with referents that they have not spoken about or interacted with, so it is worth taking this mismatch with a grain of salt.

In the unmodified conditions, there was always at least a numeric advantage for the polar interpretation, although it was only reliably preferred in the unavailable and available context conditions. Here, switching from the available to the salient comic strip context reliably affected the proportion of times each interpretation was selected, decreasing the proportion for the polar interpretation and increasing that for the scalar interpretation to the point that there was no longer a reliable preference between the two. This points to an interesting interface between the forced-choice results and the Likert results from Experiment 1; in that experiment, changing from the available comic to the salient comic with an unmodified antecedent led to significantly higher ratings for the scalar interpretation without reducing the ratings of the polar interpretation. It appears that that effect is sufficient to lead to a qualitative change in preference in a forced-choice task (from preference to indifference) even though the polar interpretation’s Likert scores were still significantly higher than those for the scalar interpretation in this condition of Experiment 1.

Furthermore, comparison of unmodified-antecedent results for Experiment 7 to those of Experiments 5 and 6 points to a familiar problem for both the strict identity account of VPE interpretation and a simple salience-based model. The identity strategy posits that in the unmodified conditions, only the polar interpretation should be available for the VPE reply. Conversely, the Experiment 5 and 6 results agreed that in the unmodified/salient condition, the specific (number information at issue) mental state was reliably more available than the indifferent (number information not at issue) mental state. Thus, the two accounts predict that in the unmodified/salient condition, one VPE interpretation should be reliably preferred over the other, namely the polar interpretation for the identity account and the scalar interpretation for the salience account. In actuality, however, the two interpretations

were approximately equally available, suggesting an interpretive account that is intermediate between these two extremes.

Finally, in the modified-antecedent conditions of Experiment 7, the scalar interpretation was always reliably preferred, as in Experiment 1. However, the results of Experiment 7 show a more nuanced effect of the context manipulation than the Likert results. In Experiment 1, the ratings for the two interpretations were roughly flat across all three modified conditions, with no reliable effect of context on the ratings. Here, there is a noticeable numeric effect of the context manipulation, with the proportion of scalar responses increasing and the proportion of polar responses decreasing as a function of number information in the comic. This effect was statistically reliable when switching from the available comic to the salient comic. This points to a possible difference between the Likert and forced-choice methodologies in their ability to detect more subtle gradations in preference when one interpretation is already significantly preferred, with the forced-choice methodology being the more sensitive.

The primary goal of this experiment was to replicate Experiment 1 with the results expressed as proportions. Since the results are highly consistent with Experiment 1, it seems safe to proceed using the result proportions as estimates of the probability of choosing one interpretation or the other in context. However, further comparison of the results of the two experiments holds the possibility of providing interesting insight into the correspondence between Likert ratings for individual interpretations and participants' likelihood of selecting a particular interpretation in a forced-choice task with multiple options, which may help guide experimental design in the future.

4.3 Probabilistic models of VPE interpretation in context

The experiments presented so far in this chapter provided two types of estimate that are important for the construction of a probabilistic model of VPE interpretation. Experiments 5 and 6 estimated the proportion of trials, in each experimental condition, that participants rated a number-modified versus -unmodified mental state as being at issue. This is intended

to estimate the “prior” probability of each meaning in the discourse context, before the utterance of a VPE construction. Experiment 7 estimated the proportion of times participants selected the polar or scalar interpretation of the VPE reply in each context, which estimates the “posterior” probability of each meaning being ascribed to the VPE utterance. With these estimates in place, they can now be used to construct and compare several probabilistic models of VPE interpretation in context.

4.3.1 *Conceptual and terminological overview*

The models considered in this chapter operate over three critical components: contexts, utterances, and meanings. In this section, a **context** will refer to the entire context of utterance prior to the utterance of the VPE reply, including both the comic strip context and the linguistic antecedent, when present. Thus, the experimental condition with a salient comic and an unmodified antecedent is a different context from the condition with a salient comic and a modified antecedent, even though they share the same comic strip context.

The **meanings** under consideration here are the polar and scalar interpretations that were discussed in the previous chapter and in Experiment 7. The **utterance** is always the VPE reply. Critically, the utterance will be taken to have different **truth values** or **compatibility ratings** with respect to a particular meaning and context, and these will vary by the theoretical interpretation strategy invoked by the different models.

The models consider a single meaning in a single context and give a formula for transforming a **prior** probability into a **posterior** probability. The prior probability is the probability that a listener thinks a particular meaning is the case in that context, and is denoted here as $P(m|c)$. For the current chapter, empirical priors are used, namely the proportions of responses in Experiments 5 and 6. In each context, participants’ empirical likelihood of selecting the indifferent mental state in these experiments is taken as the prior availability of the polar interpretation, and the empirical proportion of specific responses is taken as the prior for the scalar interpretation.

The models schematize utterance interpretation as a process of eliminating meanings that are truth conditionally incompatible with the utterance in the present context, then renormalizing the probabilities of the remaining meanings to arrive at the posterior probabilities. The general formula for this process is shown in Equation 4.1.

$$P(m|u, c) \propto \llbracket u \rrbracket^{m,c} \cdot P(m|c) \quad (4.1)$$

$\llbracket u \rrbracket^{m,c}$ represents the compatibility of the utterance with the meaning under consideration in the given context. Critically for ellipsis, this rating can depend on the context, since certain interpretation strategies require consultation of information from the broader context – for instance, the form of an antecedent VP – to determine the truth conditions of the utterance. If this rating is 0 (false) for a particular meaning in a particular context, that meaning is removed from consideration. If the rating is 1 (true), the meaning is retained as a possible interpretation, and the posterior distribution is normalized to account for the removal of any false interpretations to arrive at a proper probability distribution.

This model is directly inspired by recent work in probabilistic pragmatics, namely the Rational Speech Act framework (Bergen and Goodman, 2015; Frank and Goodman, 2012; Franke, 2009; Goodman and Stuhlmüller, 2013; Jäger, 2011; Lassiter and Goodman, 2017). Equation 4.1 is more or less identical to the *literal listener* stage of the RSA model, where a listener first consults the truth conditional content of an utterance and eliminates any literally false interpretations from consideration. Crucially, the RSA model contains multiple additional steps iterating listener and speaker reasoning about the intentions and likely actions of their interlocutor. It is not clear that VPE interpretation invokes such mechanisms, but the literal listener stage of the RSA provides a useful theoretical architecture for modeling utterance interpretation that is not necessarily pragmatically iterative.

The models presented here consider two different strategies for interpreting VPE that differ in the compatibility ratings they assign to the VPE reply, $\llbracket u \rrbracket^{m,c}$. The **identity strategy** adopts the antecedent VP from the context c wholesale in the ellipsis site. Crucially, this

means that in the context of an unmodified antecedent, the compatibility rating for the modified interpretation is 0. The **salience strategy** trivially marks all meanings as compatible with the VPE reply; the burden of deciding between the interpretations is shifted to the prior availability of the interpretations in context, as will be shown below.

Six models are considered. Two model interpretation as proceeding entirely according to the identity strategy, two exclusively according to the salience strategy, and two models hybridize these two strategies according to a proportion represented by the value β . For each of these pairs, one model considers the predictions of a straightforward implementation of Equation 4.1, while the other adds a term allowing for some random behavior in the empirical experiment (“error”, ϵ). β and ϵ are **free parameters** whose values are not known *a priori*, but are expected to have some principled impact on the transformation of the prior probability to a posterior probability.

With the model formulas and the compatibility ratings fixed, the empirical prior probabilities (proportions of responses from Experiments 5 and 6) can be substituted for $P(m|c)$ to determine the **predicted posterior probability** of a particular meaning in a particular context – that is, the model’s prediction for the probability of selecting the given interpretation in that context. The goal of the chapter is to find the model that provides the best fit between this predicted probability and the **empirical posterior probability**, which is estimated by the response proportions from Experiment 7.

Two types of analysis are considered to determine the strongest model. First, for the models that contain free parameters, a technique called **maximum likelihood estimation** is used to determine the values of the parameters that provide the “best” possible performance for that model. Of interest here is the **likelihood** of observing the empirical posterior data if the model-predicted posterior probability were actually correct. For each model, the **joint likelihood** of generating the correct proportion of empirical responses in each of the 9 conditions will be considered by multiplying the likelihoods of the 9 conditions together. Maximum likelihood estimation can then determine the parameter value(s) that maximize

the joint likelihood, indicating the instantiation of the model that is most likely to generate the entire set of empirical data.

Maximum likelihood estimation gives an idea of which model has the best “peak” performance, but it leaves open the possibility of overfitting – that is, selection of a model that happens to fit the data in one instantiation, but provides a poor fit when other parameter values are considered. This is usually a result of including too many parameters in the model, making it overly flexible and unlikely to reflect theoretically meaningful behaviors.

To account for this, another type of model comparison, the **Bayes factor**, is used. The Bayes factor is the ratio of **marginal likelihoods** for two models. The marginal likelihood is a weighted average of the likelihood of a model generating the empirical data across the entire space of possible parameter values, whereas the maximum likelihood estimate considers only one instantiation (one set of parameter values) of the model. The marginal likelihood implicitly penalizes models for overfitting, since models that provide a high peak likelihood for a small range of parameter values, but a poor fit over the bulk of the parameter space, will have low marginal likelihoods.

Each of these concepts will be discussed in more detail as they become relevant. The next subsection opens by outlining one of the simpler models, the identity model with no error term. The other models are considered in turn, followed by model comparison using the Bayes factor.

4.3.2 *Identity-only model*

The first model considered is one that schematizes VPE interpretation as proceeding under strict identity with the linguistic antecedent (when one is available). Recall that the “kernel” of each of the models in this chapter, based on the literal listener stage of the RSA model, is the statement shown in Statement 4.2.

$$P(m|u, c) \propto \llbracket u \rrbracket^{m,c} \cdot P(m|c) \tag{4.2}$$

The critical defining feature of an identity-driven model is the determination of the truth values of the elliptical utterance. The identity strategy, narrowly construed, assumes that the possible interpretations of a VPE utterance are exactly the interpretations that are available for the corresponding sentence with the antecedent VP in the position of the ellipsis site. Thus, in the current experimental paradigm, the availability of the polar and scalar interpretations can be assessed by adopting the appropriate antecedent utterance. Consider (57), which shows an unmodified antecedent and VPE reply, as well as the full-VP sentence that corresponds to the reply with the antecedent VP in situ.

(57)	I want to buy candy bars!	<i>Antecedent</i>
	We can't.	<i>VPE reply</i>
	We can't buy candy bars.	<i>Full-VP</i>

It seems clear here that the polar interpretation (*We can't buy any candy bars*) is available for the full-VP reply, while the scalar interpretation (*We can't buy five candy bars, but maybe we could buy fewer*) is not. Now, consider the same situation with a modified antecedent, as in (58).

(58)	I want to buy five candy bars!	<i>Antecedent</i>
	We can't.	<i>VPE reply</i>
	We can't buy five candy bars.	<i>Full-VP</i>

Here, it seems plausible that the father could actually intend either the polar or scalar interpretation. That is, *We can't buy five candy bars* may simply target the possibility of buying candy bars, and the father intends to communicate that they will not buy any candy bars. Alternately, the inclusion of a numeral in this utterance could be taken to target the quantity of candy bars in particular, with the father meaning to communicate that they might buy candy bars, but fewer than five.

Based on this intuition, the models in this chapter will assume that both the polar and scalar interpretations are available under identity with a modified antecedent. It is worth

considering whether this is an appropriate choice give how the utterance *We can't buy five candy bars* would be interpreted in an out-of-the-blue utterance, however. It seems clear that in a neutral context, this utterance would only be selected to convey the scalar interpretation – that is, inclusion of the numeral sends a strong signal that the quantity is at issue.

On its surface, this appears to be a pragmatic effect, while the utterance is literally compatible with buying zero candy bars. In the RSA literature, $\llbracket u \rrbracket$ in the literal listener stage considers only the semantics of the utterance, with pragmatic effects derived by the iterative reasoning stages of the model. In keeping with this approach, the present chapter will treat both the polar and scalar interpretation as available under identity with a modified antecedent, as they are semantically compatible with the full-VP sentence. However, it may be worth considering down the road whether the model should directly represent the pragmatic difficulty of deriving the polar interpretation for the full-VP sentence.

Finally, the truth values for the VPE reply with no antecedent (exophoric) must be established. A narrow identity model does not provide a clear mechanism for determining the interpretation of a VPE site that does not have an antecedent VP. However, it seems intuitively clear that, to the extent that antecedentless VPE is felicitous and interpretable, identity accounts must provide a backup mechanism that allows for such interpretation (e.g., Merchant's 2004 *do it* proposal, although Experiment 4 provided suggestive evidence that the mechanism is not inference of an elided *do that*, nor is it necessarily the case that the mechanism must provide for a one-to-one mapping between prior availability and ellipsis interpretations). Furthermore, it seems self-evident that such interpretation would necessarily be driven strictly by discourse-level effects, such as the general discourse salience of the competing interpretations.

The current chapter will gloss over the specifics of the interpretation of antecedentless VPE under an identity mechanism, and will treat VPE in cases without a VPE antecedent as trivially compatible with any interpretation. The practical effect of this in the model is that the interpretation of an antecedentless VPE site will be driven strictly by the discourse

salience of the interpretations under consideration.

Table 4.1 summarizes the compatibility ratings that will be used for the identity model. In the model specifications below, these ratings will be denoted by $\llbracket u \rrbracket_{\text{identity}}^{m,c}$, indicating the compatibility of utterance u with meaning m in context c according to an identity interpretation strategy. Recall that m denotes the interpretation or meaning under consideration, in this case polar or scalar. c denotes the context of utterance. This critically includes both the antecedent and the broader discourse context. However, only the antecedent VP is necessary to determine the compatibility ratings for the identity strategy, so the comic strip contexts are not shown in the table.

		CONTEXT, c		
		Exophoric	Unmodified	Modified
MEANING, m	Polar	1	1	1
	Scalar	1	0	1

Table 4.1: Compatibility ratings, $\llbracket u \rrbracket_{\text{identity}}^{m,c}$, for the identity interpretation strategy

Thus, the generic interpretation statement from Statement 4.2 can be rewritten for an identity-strategy model as shown in Statement 4.3.

$$P(m|u, c) \propto \llbracket u \rrbracket_{\text{identity}}^{m,c} \cdot P(m|c) \quad (4.3)$$

Intuitively, this model takes an utterance u , a meaning m , and a context c as its inputs. Meanings that are compatible with the utterance-antecedent combination have a compatibility rating of 1, so they “survive” and are possible interpretations in the posterior distribution of meanings for the utterance in context, denoted by $P(m|u, c)$. Meanings that are incompatible with the utterance-antecedent combination have compatibility ratings of 0, which automatically transforms their availability in the posterior distribution to 0.

Thus, the model states that the probability of concluding a particular meaning from the VPE reply in a particular context is proportional to the meaning’s prior probability in the context, subject to the constraint that all literally incompatible meanings have been

removed. In the context of the current paradigm, this means that in the exophoric and modified conditions, the probability of the polar versus scalar interpretation is proportional to their prior availability in a particular context, whereas only the polar interpretation should be available with a modified antecedent because the scalar interpretation has a compatibility rating of 0.

The only remaining modification that is necessary to Statement 4.3 is to transform the proportionality statement, denoted by $\propto$, into a concrete calculation. As it stands now, the posterior distribution that would be generated by Statement 4.3 is not necessarily a proper probability distribution in that the posterior probabilities need not add up to 1. To understand this, consider the set of all possible meanings, denoted by M , containing meanings $m_1, m_2, \dots m_k$. The prior distribution of possible meanings in a given context is a proper probability distribution. However, the model “transforms” this distribution by eliminating meanings that are incompatible with the chosen combination of utterance and antecedent. If any meaning is eliminated by this process, then the output distribution will not sum to 1 (since the prior probabilities of compatible meanings are merely multiplied by 1).

The model addresses this by normalizing the values output by the right side of Statement 4.3. This is done by dividing the output by the sum of all of the outputs for individual meanings in the set of possible meanings M . The practical effect of this is that the raw numbers output by Statement 4.3 are transformed so they indicate their relative “strength,” as a proportion, relative to only the meanings that “survived” the compatibility ratings. In other words, the sum represents the bulk of the prior probability distribution that “survives” the transformation, and dividing individual outputs by this sum changes the raw output to a proportion of this bulk that corresponds to the individual meaning in question. These proportions can be interpreted as probabilities, since the output proportions for all possible meanings will now sum to 1.

Thus, the final identity model, including a denominator for normalizing the output values,

is shown in Equation 4.4. In the denominator, m' denotes any arbitrary meaning m in the set of possible meanings M , and the summation indicated by Σ represents the operation of adding up all of the outputs of the indicated expression for all possible values of m' . Note that for the purposes of modeling in this chapter, M will be taken to consist of only the polar and scalar interpretations for the sake of simplicity, although of course in a naturalistic context it is likely that the situation “on the ground” is considerably messier and additional meanings may be under consideration.

$$P(m|u, c) = \frac{\llbracket u \rrbracket_{\text{identity}}^{m,c} \cdot P(m|c)}{\sum_{m' \in M} \llbracket u \rrbracket_{\text{identity}}^{m',c} \cdot P(m'|c)} \quad (4.4)$$

With the model determined, it can now be used to transform the empirical prior probabilities from Experiments 5 and 6 into a posterior distribution describing the probability of arriving at a particular interpretation for the VPE reply. This is done by using Equation 4.4 to calculate the probabilities of the polar and scalar VPE interpretations in each context using the prior probabilities of the indifferent and specific interpretations in the same context and the utterance compatibility ratings as inputs.

Recall that the probability of the indifferent mental state is being used as the prior probability of the polar interpretation, and the probability of the scalar interpretation is treated as the prior probability of the scalar interpretation in context. Thus, when calculating the posterior probability of the polar interpretation, the polar compatibility ratings and the indifferent prior are used in the numerator of Equation 4.4, and when calculating the scalar posterior, the scalar compatibility ratings and specific prior are used in the numerator.

Sample calculations demonstrating the use of each model to transform the empirical priors to a probability distribution are shown in Appendix H. For the analysis presented below, the empirical priors from Experiment 5 were used. The three-alternative task was impressionistically more natural, since there are some conditions, particularly the unavailable/exophoric condition, where the zero mental state was expected to be available, and the

two-alternative Experiment 6 left this option out. The three-alternative priors also yield an impressionistically better fit to the empirical posterior data, although the results of the modeling analysis with the two-alternative priors are not qualitatively different.

To transform the Experiment 5 results to proportions for use as inputs to the models, the zero interpretations were discarded and the proportions of polar and scalar responses in each context were recalculated after this omission. It is worth noting that the models are actually theoretically capable of handling more than two possible interpretations. The zero mental state was excluded because it was not clear that the zero mental state as a prior maps onto either of the forced-choice interpretations for the VPE reply. That is, if the discourse is roughly schematized by the antecedent *I don't want to buy any candy bars* and the reply *We can't*, it is not clear that either the polar or the scalar interpretation should be available.

Figure 4.7 plots the predicted probabilities, according to the identity model without error, of concluding the polar and scalar interpretations in each context against the empirical interpretation proportions observed in Experiment 7. In this plot, predicted probabilities according to the model are plotted in gray, while the empirical proportions observed in Experiment 7 are plotted in black. Circular points show the probability or proportion for the polar interpretation, while triangular points show the probability or proportion for the scalar interpretation. Thus, a condition where the black and gray circles and the black and gray triangles are close to one another, respectively, shows a strong fit between the model predictions and the empirical data. A condition where points of the same shape and different colors are far from one another shows a poor model fit.

A visual inspection of this plot indicates a relatively close mapping between the model-predicted interpretations and the empirical results in the exophoric and modified-antecedent conditions, but a major mismatch in the unmodified-antecedent conditions. In the exophoric and modified conditions, VPE interpretation is assumed in the identity model to be influenced by the prior availability of each interpretation in the given context, since the VPE utterance interpreted according to the identity strategy is semantically compatible with both

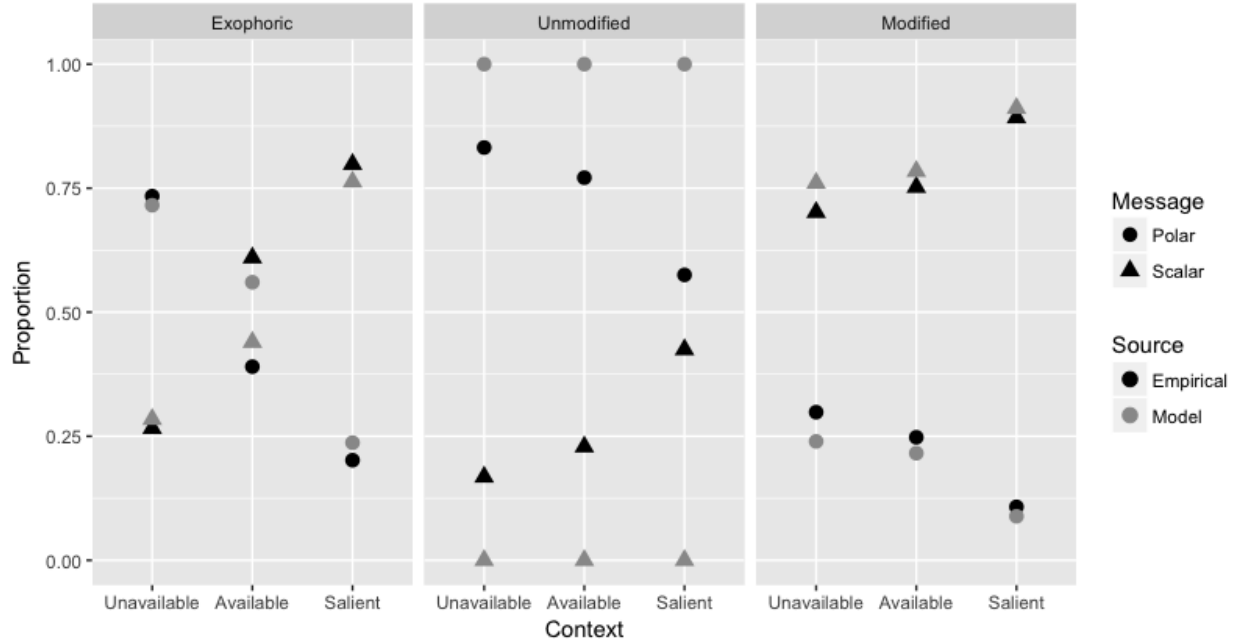


Figure 4.7: Predicted posteriors for identity model (no error) versus empirical posteriors.

the polar and scalar interpretations with no antecedent or with a modified antecedent.

However, the identity strategy bans the scalar interpretation when the antecedent is unmodified. As a result, the simple identity model predicts a probability of 1 of concluding the polar interpretation and a probability of 0 of selecting the scalar interpretation in these conditions. There is a clear mismatch between this prediction and the empirical data. While the polar interpretation is strongly preferred in the unavailable/unmodified and available/unmodified conditions, it was not selected 100% of the time. More importantly, the polar interpretation was not even significantly preferred over the scalar interpretation in the empirical results, with the polar interpretation selected closer to 60% of the time.

This mismatch is potentially problematic for a strict identity account of VPE interpretation. As discussed in the previous chapter, there appears to be a clear effect of discourse on the interpretation of VPE even in the context of an unmodified antecedent, where according to the identity strategy broader context information should be ignored. The model discussed immediately below explores an option for improving the fit of the identity-strategy model without resorting to the inclusion of discourse information.

4.3.3 *Identity model with error*

As discussed above, the simple identity model yielded highly flawed predictions relative to the empirical data in the unmodified-antecedent conditions. While there appears to be a reliable effect of discourse in these conditions, some features of the empirical data suggest a second possible approach to modeling the identity strategy.

It is worth noting that across all of the conditions of Experiment 7, the “ceiling” of proportions is not 1. Even in the condition with the most extreme gap in preference between the two interpretations, the salient/modified condition, the preferred scalar interpretation was selected only about 90% of the time. This is perhaps unsurprising for two reasons. First, as a general observation on utterance interpretation, it seems likely that things will “go wrong” some of the time, with listeners misinterpreting the communicative intent of the speaker, forgetting certain features of the discourse context, misunderstanding the utterance, and so on. Second, in the specific task for Experiment 7, it would not be surprising to observe that some trials reflect inattention or random guessing from the AMT participants, even after the data were trimmed to avoid including such trials.

In light of this, it is worth considering a model that allows for a certain amount of random behavior on the part of interpreters. This will be referred to here as “error” – essentially, this models listeners as guessing at random from a list of possible interpretations on some proportion of interpretation trials rather than using the theoretically motivated interpretation strategy. The error discussed here will gloss over the distinction between error that is likely to occur in naturalistic interpretation due to signal noise, forgetting features of the prior discourse, etc., and error that is specific to the AMT task.

Equation 4.5 shows the identity-strategy model augmented with a free parameter allowing for occasional random guessing. The parameter ϵ represents the probability of a listener guessing at random from the possible interpretations.

$$P(m|u, c) = (1 - \epsilon) \frac{\llbracket u \rrbracket_{\text{identity}}^{m,c} \cdot P(m|c)}{\sum_{m' \in M} \llbracket u \rrbracket_{\text{identity}}^{m',c} \cdot P(m'|c)} + \epsilon \frac{1}{|M|} \quad (4.5)$$

In this equation, the original identity-strategy calculation from Equation 4.4 is multiplied by $1 - \epsilon$, schematizing that listeners have a probability of $1 - \epsilon$ of using the identity strategy to interpret a VPE utterance. The second term models random selections. Recall that the equation yields the probability of selecting a particular meaning for a given utterance in a particular context. $\frac{1}{|M|}$ in this term represents the probability of selecting that meaning at random from the set of possible meanings. If there are two possible meanings, the probability of selecting the meaning is $\frac{1}{2}$, if there are three, $\frac{1}{3}$, and so on. In this case, since the set of possible meanings consists of the polar and scalar interpretations, the value of $\frac{1}{|M|}$ is $\frac{1}{2}$. This fraction is multiplied by ϵ , the probability of using the random-selection strategy rather than the identity strategy.

This model introduces a free parameter, ϵ , so the predicted probability for a particular interpretation in a particular context depends on the value selected for this parameter. The minimum theoretically motivated value for the free parameter is 0, since ϵ represents a probability or a proportion of trials. The instantiation of the model with $\epsilon = 0$ is identical to the identity model without error presented above, since with an ϵ of 0 the second term in Equation 4.5 is eliminated.

As ϵ increases, the proportion of the posterior probability that is determined by the random-selection strategy increases, and the proportion determined by the identity interpretation strategy decreases. At the maximum possible value of ϵ , which is 1, the predicted posterior probability for each interpretation in each condition is $\frac{1}{2}$. This is because $1 - \epsilon$ is 0, so the identity interpretation term drops out, and the model simplifies to $\frac{1}{|M|} = \frac{1}{2}$. For intermediate values, as ϵ increases, the predicted probabilities move closer to $\frac{1}{2}$ relative to the probabilities predicted by the identity model without error (with $\epsilon = 0$).

Intuitively, there should be some ϵ value between 0 – which gives the incorrect categorical

prediction in the unmodified conditions – and 1 – which predicts a probability of $\frac{1}{2}$ across the board – that maximizes the fit between the predicted probabilities for each interpretation and the actual empirical proportions observed in Experiment 7. This chapter considers the “maximum likelihood” instantiation of each model with free parameters, including the current identity-with-error model, in order to consider the “best” possible performance of each model as a means of describing the interpretation process.

Since there are two possible interpretations in the current application, the individual conditions can be thought of in terms of the binomial distribution. On the one hand, the output of the model is a probability of selecting, say, the polar interpretation in a particular context. On the other hand, Experiment 7 consists of some number of trials n in each condition, in which a certain number of responses, x , were polar rather than scalar. Thus, the *likelihood* of a particular instantiation of the model generating the experimental data in a particular condition is the probability, according to the binomial distribution, of selecting the polar interpretation in x out of n trials when the probability of selecting the polar interpretation on any particular trial is the model output. This can be calculated in R using the `dbinom` function from the package `stats`.

In turn, the likelihood of the model generating the experimental data in all nine conditions is simply the product of its likelihood of generating the empirical data in each of the conditions individually. This value is likely to be infinitesimal, but it is calculable.

Thus, identifying the maximum likelihood instantiation of a particular model means maximizing this joint likelihood, the product of the likelihoods in all of the conditions. For the current chapter, this was accomplished using the maximum likelihood estimation function `mle` from the R package `stats4`. Given a particular model and a means of calculating its likelihood of generating some set of data, this function considers different parameter values within a specified possibility space and determines the value that maximizes the likelihood. (In this case, it was used to minimize the sum of the negative log likelihoods in each condition, which is equivalent to maximizing the product of the individual-condition likelihoods).

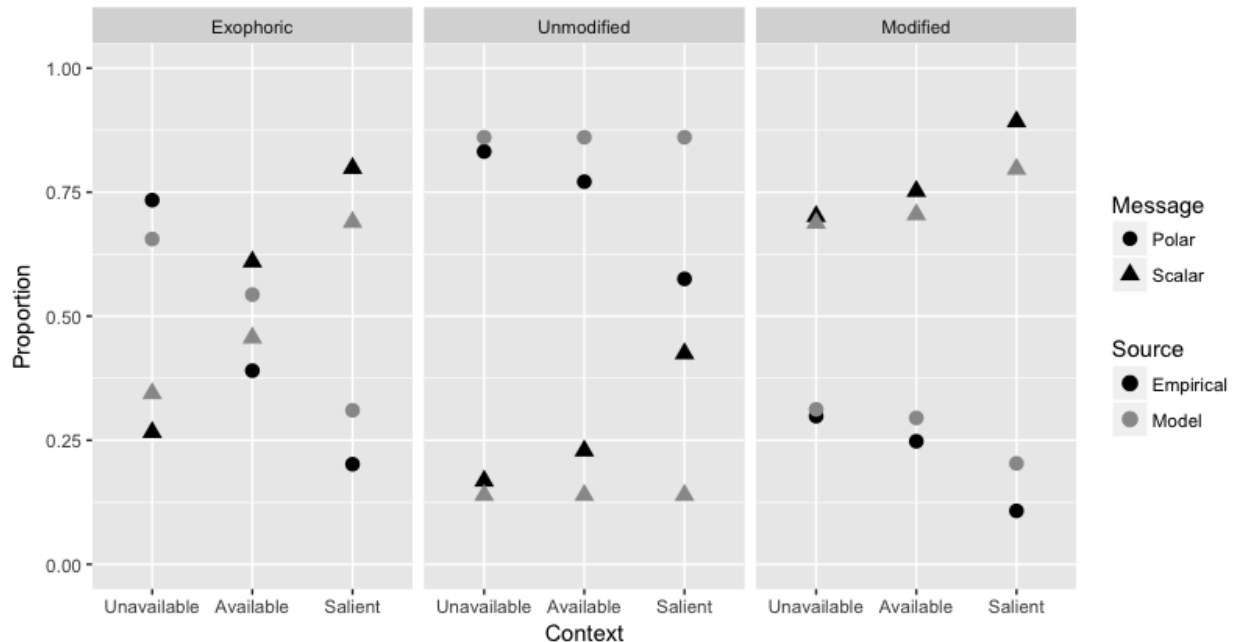


Figure 4.8: Predicted posteriors for identity model with error versus empirical posteriors. Maximum likelihood instantiation of model with $\epsilon \approx 0.279$.

For the identity-with-error model, the `mle` package was used with algorithm L-BFGS-B to restrict the possible values of ϵ to be between 0 and 1, inclusive. The output indicated that the maximum likelihood instantiation of this model has an ϵ value of approximately 0.279. Figure 4.8 plots the predictions of the identity-with-error model with this parameter value against the empirical proportions from Experiment 7.

A visual inspection of this figure reveals several problems with the fit between the model predictions and the empirical data. First, although the predictions in the modified conditions are much closer to the empirical data than the predictions of the identity model without error, the fit in these conditions still leaves something to be desired. While the predicted and empirical values are similar in the unavailable/unmodified condition, they are further apart in the available/unmodified condition, and quite far apart in the salient/unmodified condition.

Intuitively, this is not surprising. The identity-with-error model uses a one-size-fits-all strategy to derive antecedent-nonidentical interpretations in the unmodified condition, posit-

ing that any selection of the scalar interpretation with the antecedent VP *buy candy bars* is due to listeners guessing at random between interpretations. Adjusting the value of ϵ relative to 0 (that is, the identity model without error) can only move the predicted probabilities in all three unmodified conditions in unison. In other words, the model is incapable of representing the most striking empirical fact of interpretation in the unmodified condition, which is the fact that different discourse contexts lead to reliably different proportions for the polar and scalar interpretations.

Second, the fit between the predicted and empirical data is notably worse in the maximum-likelihood instantiation of the identity-with-error model compared to the identity model without error in several conditions, including the unavailable/exophoric, salient/exophoric, and salient/modified conditions. In order to provide any semblance of fit in the unmodified conditions, the model must resort to a very high value of ϵ to move the predicted polar probability from 1 to something relatively closer to 0.5. However, a consequence of this is that the predicted probabilities in the remaining six conditions must move closer to 0.5, as well. As a result, in conditions where there was already a strong mapping between the predicted and empirical probabilities, the fit actually gets worse as the predicted probability is moved so dramatically in the direction of 0.5.

Relatedly, it is worth considering the real-world meaning of the proposed ϵ value of 0.279. This means that participants are modeled as guessing between the polar and scalar interpretations at random 27.9% of the time, and using the theoretically motivated identity strategy to interpret VPE only 72.1% of the time. While the ϵ parameter is included specifically to account for the possibility of random behavior, it would be alarming to discover that listeners behave randomly when interpreting utterances nearly 30% of the time, even for potentially difficult-to-interpret utterances like VPE sentences in complex discourse contexts.

In sum, a probabilistic model based on a strict identity strategy of VPE interpretation, even when augmented with the possibility of random guessing and other noise in the experimental data, provides a poor fit to the empirical data from Experiment 7. In addition,

the maximum likelihood instantiation of the identity-with-error model posits an implausibly high estimate of how often participants guess at random during utterance interpretation, since it uses the error term as a brute force means of deriving antecedent-nonidentical interpretations in the unmodified conditions. These weaknesses are discussed further in the model comparison and discussion below.

4.3.4 *Saliency-only model*

As an alternative to the identity models discussed above, this subsection considers a model that schematizes VPE interpretation as proceeding entirely according to the prior discourse saliency of the competing propositions. As a starting point, the model shown in Equation 4.6 is identical to the identity model without error, except that it replaces the compatibility ratings from the identity strategy, $\llbracket u \rrbracket_{\text{identity}}^{m,c}$, with a set of compatibility ratings corresponding to the discourse interpretation strategy, $\llbracket u \rrbracket_{\text{saliency}}^{m,c}$.

$$P(m|u, c) = \frac{\llbracket u \rrbracket_{\text{saliency}}^{m,c} \cdot P(m|c)}{\sum_{m' \in M} \llbracket u \rrbracket_{\text{saliency}}^{m',c} \cdot P(m'|c)} \quad (4.6)$$

For the current modeling purposes, the discourse strategy will be taken to be one that marks any interpretation as trivially compatible with a VPE utterance. That is, there is no firm grammatical restriction on the interpretations that can arise for VPE; interpretation is instead guided entirely by independent features of the discourse, such as the discourse availability of the competing interpretations. This is clearly an oversimplification of actual interpretation, where some meanings will necessarily be incompatible with particular elliptical constructions even if the discourse interpretation strategy is correct. In this application, however, it seems reasonable to posit that both the polar and scalar interpretations are compatible with the form of the VPE reply.

Thus, Table 4.2 summarizes the compatibility ratings that will be used in the saliency model without error.

		CONTEXT, c		
		Exophoric	Unmodified	Modified
MEANING, m	Polar	1	1	1
	Scalar	1	1	1

Table 4.2: Compatibility ratings, $\llbracket u \rrbracket_{\text{salience}}^{m,c}$, for the discourse interpretation strategy

These compatibility ratings allow for some notational simplification of the model shown in Equation 4.6. Since $\llbracket u \rrbracket_{\text{salience}}^{m,c}$ is always 1, the numerator of the model simplifies to $P(m|c)$, the prior probability of the meaning under consideration in the given context. Since no interpretations are eliminated from consideration by having a compatibility rating of 0, the sum in the denominator is always 1. Thus, the model simplifies to Equation 4.7.

$$P(m|u, c) = P(m|c) \quad (4.7)$$

Thus, the discourse model without error simply predicts that the posterior probability of selecting a particular interpretation is equal to that interpretation’s prior probability in the given context. This is probably the simplest possible “discourse” model of VPE interpretation. It is important to note here that this corresponds to the “simple” discourse model as discussed in the previous two chapters. Notably, this model does not correspond to certain discourse-driven models that have been proposed in the literature. For instance, Hardt (1993) treats VPE as a null proform, but the selection of an antecedent meaning is guided by additional heuristics beyond just the salience of a particular property in the discourse, such as subject coreference. Nevertheless, it seems worthwhile to consider the simplest model and systematically complexify it. This salience-only model may also have a precursor in Miller and Pullum’s 2013 account of VPE licensing. They propose that VPE is licensed by a salient propositional contrast in the discourse context, which may invite an analysis of interpretation in which it consists simply of identifying the contrast that was most available and selecting the proposition that matches the form of the ellipsis site.

Figure 4.9 shows the predicted posterior probabilities according to the salience model

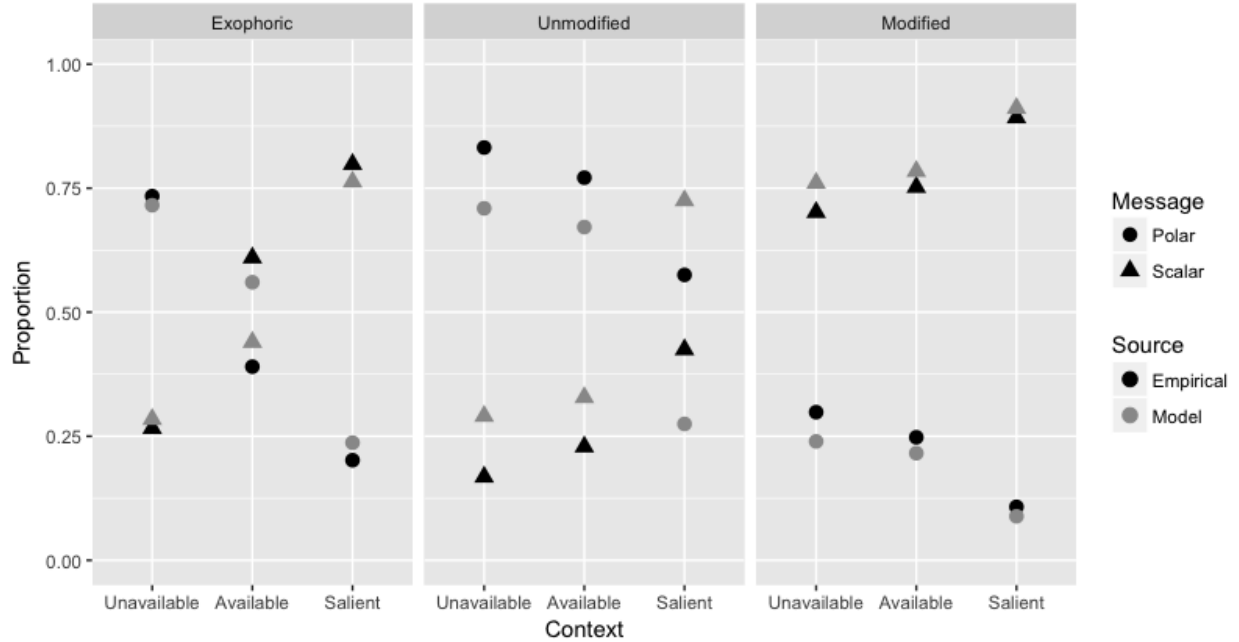


Figure 4.9: Predicted posteriors for salience model (no error) versus empirical posteriors.

without error plotted against the empirical posteriors from Experiment 7. Note that the predicted posteriors are identical to the empirical priors elicited in Experiment 5.

A visual inspection of this plot reveals many of the same shortcomings that have already been noted for a salience-only model of VPE interpretation. Most strikingly, the salience strategy systematically overestimates the availability of the scalar interpretation in the unmodified-antecedent conditions. This is particularly striking in the salient/unmodified condition, where the scalar interpretation has a reliably higher prior probability (specific mental state) than the polar interpretation (indifferent mental state), yet in actual VPE interpretation listeners are roughly indifferent between the polar and scalar interpretations. As has been discussed previously, this suggests that interpretation is actually constrained by the form of the linguistic antecedent, separate from its contribution to salience, as the identity strategy predicts that the scalar interpretation should be unavailable (or, on a more permissive reading, less available) when the antecedent is not modified.

4.3.5 *Saliency model with error*

The next model considered augments the previous discourse-driven model with an error term of the same form as the one from the identity-with-error model. This term provides models with some theoretically motivated flexibility to escape the rigid predictions made by the posited interpretation strategy, so it seems appropriate to afford the same courtesy to the saliency-driven model that was given to the identity-driven model.

Equation 4.8 shows the discourse model with error. Recall that $P(m|c)$ is a mathematical simplification of the theoretically motivated discourse interpretation strategy, a consequence of the fact that every interpretation is marked as compatible with VPE according to the discourse strategy.

$$P(m|u, c) = (1 - \epsilon) \cdot P(m|c) + \epsilon \frac{1}{|M|} \quad (4.8)$$

This model predicts that listeners have a probability of $1 - \epsilon$ of faithfully executing the discourse-driven strategy, that is, that the interpretation under consideration will be selected in proportion to its prior availability in the discourse context. There is a probability of ϵ that the listener will instead select an interpretation at random; in this case, $\frac{1}{|M|} = \frac{1}{2}$.

The maximum likelihood instantiation of this model was determined using the same method as for the identity-with-error model, considering possible values of ϵ from 0 to 1, inclusive. The joint likelihood of the model generating the empirical posterior data in each condition was maximized with an ϵ value of about 0.084. The predicted posteriors from this instantiation of the model are plotted against the empirical posteriors in Figure 4.10.

The likelihood was maximized with an ϵ value of 0.084, corresponding to real-world behavior of guessing at random about 8.4% of the time. This seems like a reasonable value, at least for the AMT task, which could be expected to include some more guesses than would be typical in a naturalistic conversation.

However, the addition of this guessing behavior does not seem impressionistically to have

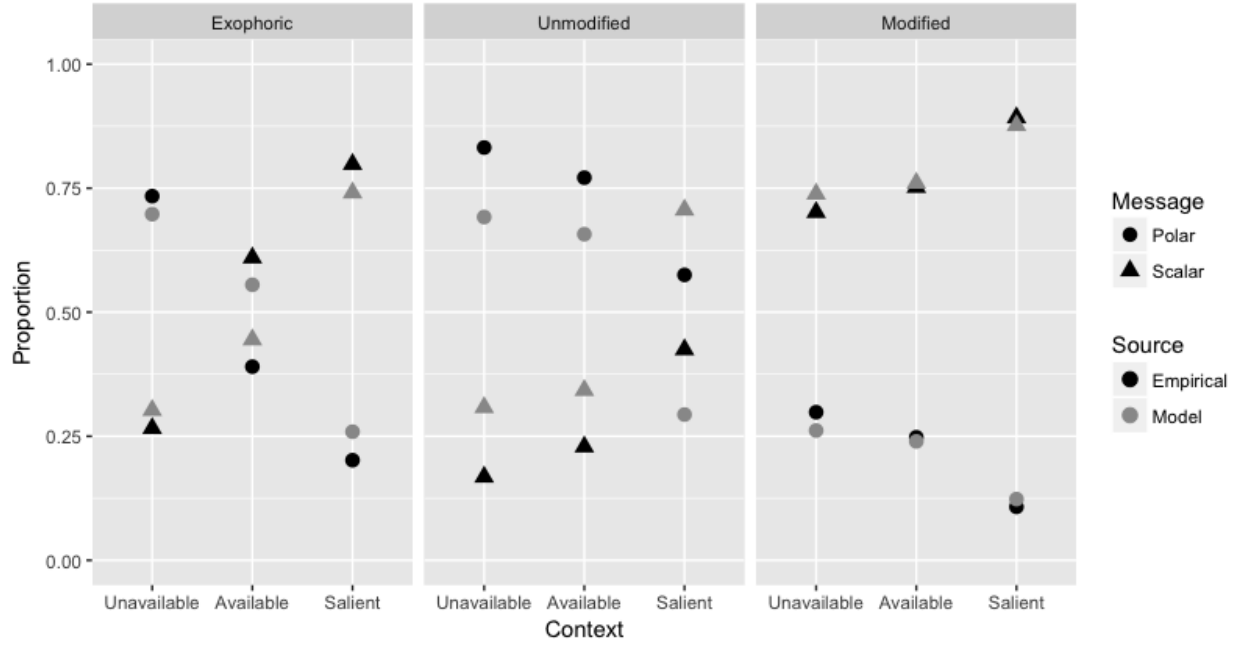


Figure 4.10: Predicted posteriors for salience model with error versus empirical posteriors. Maximum likelihood instantiation of model with $\epsilon \approx 0.084$.

substantially improved the fit between the predicted and empirical posterior probabilities compared to the salience model without error. Recall that the non-error models are equivalent to the error models with an ϵ value of 0. Increasing the value of ϵ from 0 moves the predicted interpretations closer to 0.5, since the value of $\frac{1}{|M|}$ is $\frac{1}{2}$.

In this case, increasing ϵ improves the fit in cases where the predicted probability was further from 0.5 than the empirical proportion was, including the salient/unmodified conditions and all three modified-antecedent conditions. However, this comes at the expense of decreased fit in the conditions where the predicted probability was already closer to 0.5 than the empirical probability was, including the unavailable/exophoric, available/exophoric, unavailable/unmodified, and available/unmodified conditions. Thus, the utility of the error term as a means of improving fit is limited, and the likelihood happens to be maximized at a low value of ϵ .

As a result, even the maximized likelihood instantiation of the salience-with-error model suffers from many of the same limitations as the salience model without error. Chief among

these is the fact that it overestimates the availability of the scalar VPE interpretation in the unmodified-antecedent conditions, where interpretation actually appears to be constrained by identity with the antecedent, which does not make number information available.

4.3.6 Hybrid model (no error)

The next model considered schematizes VPE interpretation as a hybrid of the identity and discourse interpretation strategies. Both of the “simple” models discussed above suffer from limitations in their ability to generate the empirical VPE interpretation data. The identity strategy fails to account for the effect of the comic strip context manipulation on the distribution of interpretations in the unmodified-antecedent conditions, where it predicts that the scalar interpretation should be unavailable. By contrast, the discourse strategy, in its simplest form considering only prior salience, fails to account for the constraints the form of the antecedent places on interpretation, with the availability of the scalar interpretation in the unmodified-antecedent conditions substantially lower under ellipsis than in the prior context at large.

Intuitively, these two issues represent two sides of the same coin in the unmodified-antecedent conditions, and could be addressed by “averaging out” the predictions of the two-interpretation strategy. Equation 4.9 shows a model that takes this approach.

$$P(m|u, c) = \beta \cdot \frac{\llbracket u \rrbracket_{\text{identity}}^{m,c} \cdot P(m|c)}{\sum_{m' \in M} \llbracket u \rrbracket_{\text{identity}}^{m',c} \cdot P(m'|c)} + (1 - \beta) \cdot P(m|c) \quad (4.9)$$

This model introduces a new free parameter, β , schematizing the combination of both the identity and discourse strategies. In essence, a proportion of β of the predicted posterior probability is determined by the identity strategy; note that the factor multiplied by β is the identity model without error. Likewise, a proportion of $1 - \beta$ of the posterior probability is determined by the salience strategy. Recall that $P(m|c)$, which is multiplied by $1 - \beta$, is the

simplified form of the salience model without error.

There are several possible real-world interpretations of β that will largely be glossed over for the time being. One possibility is that for any given instance of utterance interpretation, a listener has a probability of β of faithfully executing the identity strategy, and a probability of $1 - \beta$ of faithfully executing the discourse strategy. Another possibility is that β actually represents a weighting parameter that listeners use to combine the predictions of the two strategies to determine their confidence in a particular interpretation on all trials. These possibilities are taken up in more detail in the discussion below.

Since β represents a proportion, the values it can theoretically take on range from 0 to 1, inclusive. Note that when β is 1, the model is equivalent to the identity model without error, and when it is 0, the model is equivalent to the salience model without error. Using the same technique as for the error models, the maximum likelihood instantiation of this model was estimated. Note that when β is 1, the model’s likelihood of generating the experimental posterior data is 0, because the model predicts that the polar interpretation should be selected 100% of the time in the unmodified-antecedent conditions.

Because of this, the maximum likelihood estimation was constrained to consider β values from 0 to 0.99, inclusive. Intuitively, it seems safe to exclude this upper range of β values; given the magnitude of the predicted-empirical mismatch for the identity model in the unmodified conditions, it seems unlikely that the value of β that would “save” the model is one that adjusts the interpretation toward the discourse strategy by less than one percent.

The analysis showed that the likelihood of this hybrid model generating the experimental data is maximized with a β value of approximately 0.594. The predictions of this instantiation of the model are plotted against the empirical proportions in Figure 4.11.

The fit between the posterior probabilities predicted by the hybrid model and the empirical data from Experiment 5 is striking when compared to the models considered so far. Perhaps most notable is the performance of the model in the unmodified conditions. The model correctly predicts a strong preference for the polar interpretation in the unavailable

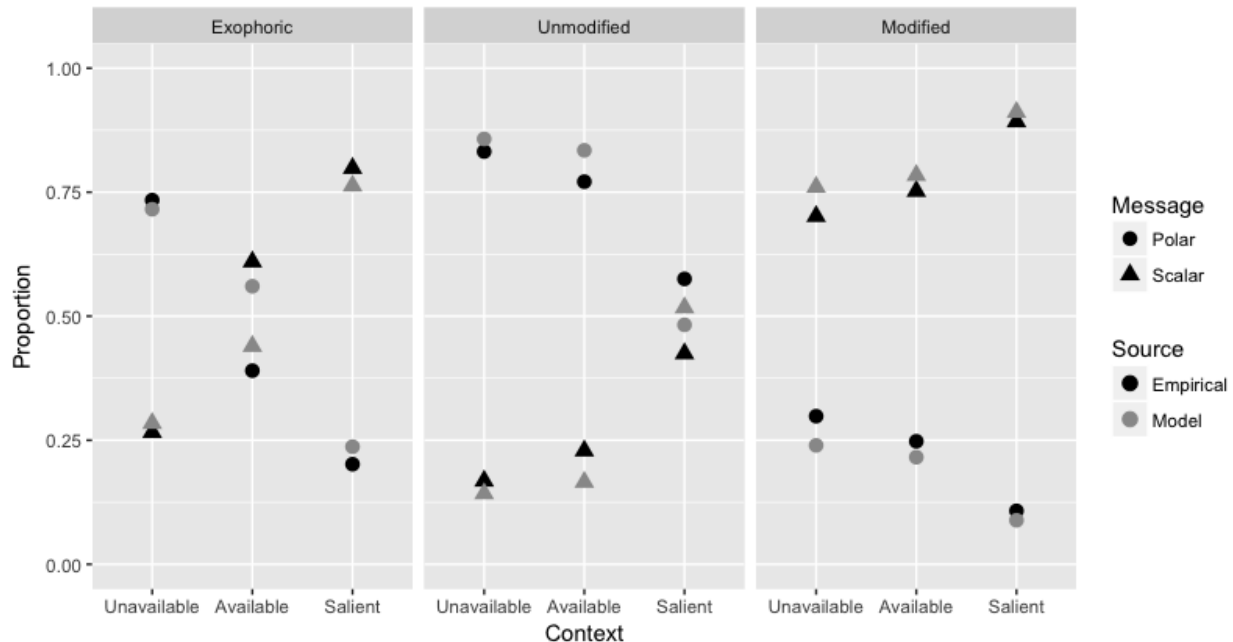


Figure 4.11: Predicted posteriors for hybrid model (no error) versus empirical posteriors. Maximum likelihood instantiation of model with $\beta \approx 0.594$.

and available contexts, but approximate indifference between the polar and scalar interpretations with a salient context. As expected, combining the predictions of the two interpretation strategies in the correct proportion appears to correspond well to the empirical facts of interpretation in the unmodified conditions; the availability of the scalar interpretation increases as a function of number information in the comic strip context, but is blocked from achieving a proportion of selections equal to its prior availability in any of the three contexts by the fact that it is formally ruled out by identity with the unmodified antecedent.

The most apparent predicted-empirical mismatch for this model is the available/exophoric condition, which has largely been ignored until now. None of the models considered in this chapter perform particularly well in this condition. This is an artifact of the numerical preference for the polar prior (indifferent mental state) in Experiment 5 on the one hand, and the numerical preference for the scalar interpretation under VPE in Experiment 7 on the other. In the models with no error term, the model-predicted posterior in this condition is equal to the prior, so the mismatched observed between the results in the available/exophoric condi-

tion is preserved. Even in the models with error terms, the best possible fit in this condition is for the model-predicted probabilities for the polar and scalar interpretations to be 0.5. This is because, intuitively, the model prediction “starts” on the “wrong” side of 0.5 relative to the empirical posterior, and can only be moved closer to 0.5, but not past it, by increasing the value of ϵ .

This issue appears to stem from the fact that participants in this condition were roughly indifferent between the two mental states in Experiment 5 and the two VPE interpretations in Experiment 7, with the corresponding mental state-VPE interpretation pairs falling on opposite sides of 0.5 simply by chance. For this reason, poor fit in this condition will not be treated as a problem with the underlying theoretical architecture of any of the individual models.

4.3.7 Hybrid model with error

The final model considered in this chapter augments the hybrid model discussed immediately above with an error term. It is not immediately clear from Figure 4.11 that this will substantially improve the fit. Recall that increasing error from 0 shifts the model-predicted interpretations closer to 0.5. For the instantiation of the hybrid model shown in Figure 4.11, this would improve the fit somewhat in seven conditions, but make it worse in the remaining two, where the model predictions are already closer to 0.5 than the empirical proportions are. Furthermore, the utility of shifting toward 0.5 is limited in each condition, either because the model prediction will get asymptotically close to 0.5, but remain on the opposite “side” from the empirical proportion, or because the model prediction will ultimately “pass” the empirical proportion and further shifts toward 0.5 would worsen fit.

Nevertheless, it is worthwhile to explore what effect including an error term in the hybrid model has, if simply for the sake of consistency across the three different types of models (identity, salience, and hybrid). Equation 4.10 shows the specification of the hybrid model with the error term incorporated.

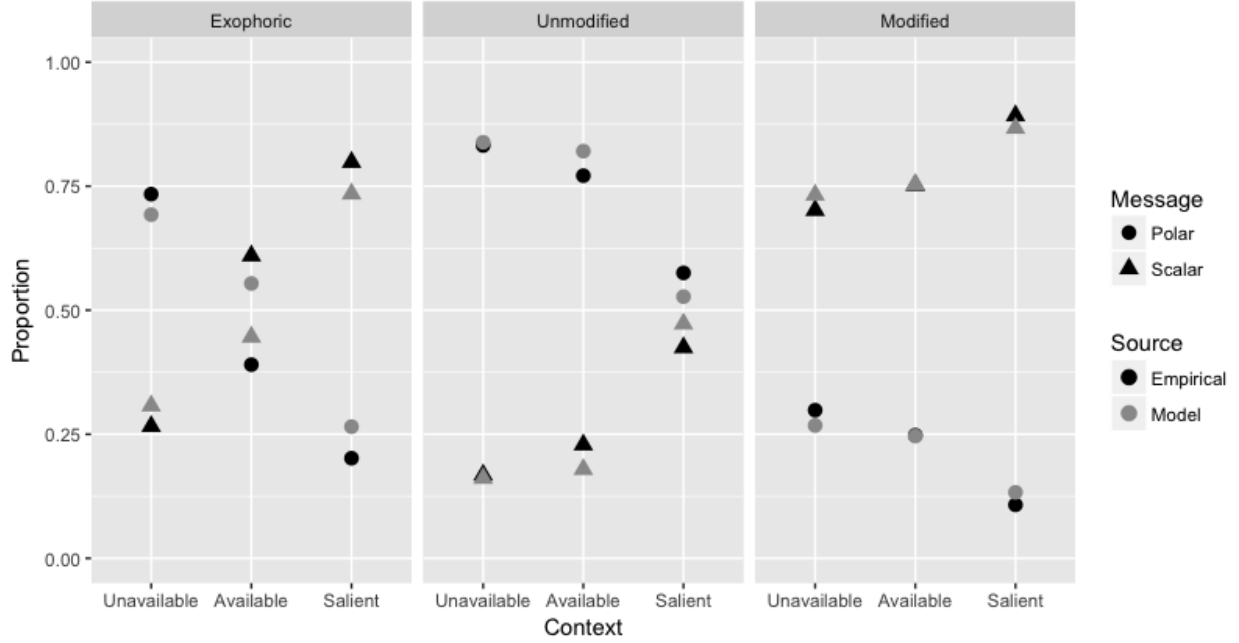


Figure 4.12: Predicted posteriors for hybrid model with error versus empirical posteriors. Maximum likelihood instantiation of model with $\beta \approx 0.665$ and $\epsilon \approx 0.107$.

$$P(m|u, c) = (1 - \epsilon) \left(\beta \cdot \frac{\llbracket u \rrbracket_{\text{identity}}^{m,c} \cdot P(m|c)}{\sum_{m' \in M} \llbracket u \rrbracket_{\text{identity}}^{m',c} \cdot P(m'|c)} + (1 - \beta) \cdot P(m|c) \right) + \epsilon \cdot \frac{1}{|M|} \quad (4.10)$$

According to this model, there is a probability of $1 - \epsilon$ that the interpretation is determined by the hybrid model without error described immediately above, which is identical to the factor multiplied by $(1 - \epsilon)$ in the equation. As in the other error models, there is a probability of ϵ that the interpretation will instead be selected at random.

As before, the `mle` function in R was used to estimate the maximum likelihood instantiation of the model taking into account various values for both β and ϵ . The possible values for β were restricted to 0 to 0.99, inclusive, and the possible values for ϵ were restricted to 0 to 1, inclusive. The maximum likelihood instantiation of the model had $\beta \approx 0.665$ and $\epsilon \approx 0.107$. The predicted probabilities according to this instantiation are plotted against the empirical proportions in Figure 4.12.

As expected, the addition of an error term did not yield an immediately apparent improvement in fit compared to the hybrid model without error. The fit is extremely good in some conditions, such as the unavailable/unmodified and available/modified conditions. However, this comes at the expense of fit in other conditions, like unavailable/exophoric and salient/exophoric, where the model prediction in the hybrid model without error was already closer to 0.5 than the empirical proportion was. The adjustment of the model predictions toward 0.5 also “overshoots” the empirical proportion in the salient/modified condition, where the model predictions were further from 0.5 than the empirical proportions were in the previous model, but are closer to 0.5 in the current model.

4.3.8 Model comparison

With the different models specified, they can now be compared to one another to determine which one provides the best fit to the empirical data from Experiment 7. It may seem obvious from the plots accompanying each model that the hybrid models provide a substantially better fit than any of the identity- or salience-driven models. Indeed, when considering the maximum likelihood instantiation of each model, the joint likelihoods for the two hybrid models are much higher than those for any of the other four models.

However, a major consideration when evaluating the utility of a particular probabilistic model is checking the fit across the entire possibility space for any free parameters rather than simply examining the single parameter value that provides the best fit. This is important for several reasons. First, a particular model might “accidentally” yield a strong fit for a narrow range of parameter values, but miss the empirical data badly for other values. This model might be dispreferred to a model that provides a worse fit in its “best” instantiation, but a better fit across the entire parameter space, as the latter could be argued to provide a better model of the real processes underlying the generation of the empirical data in that it doesn’t “guess” the empirical data as a random one-off observation.

The second major consideration relates to the complexity of the models under consider-

ation in terms of the number of free parameters. In general, it will be easier for a model with more free parameters to generate a given set of empirical data than one with fewer parameters. To illustrate this concretely, consider some pairs of proposed models of VPE interpretation.

The identity model without error is locked into the predicted probabilities that result from transforming the empirical priors according to the utterance-meaning compatibility ratings. It has no free parameters, so there is no variation in the output probabilities. By contrast, the identity-with-error model contains a free parameter that allows for adjustment of the identity-strategy outputs. Compared to the identity model without error, the with-error model has the option of increasing the value of ϵ , moving each of the predicted probabilities closer to 0.5. Adjusting the specific value of the parameter changes the strength of predicted-empirical fit in each condition, and by extension, the joint likelihood across the entire experiment. Intuitively, it is highly likely that a model with one free parameter will have at least one instantiation with a higher likelihood than a model with zero parameters, which is what is observed with the identity models.

Similarly, consider the hybrid model without error, which has one free parameter, and the hybrid model with error, which has two. The hybrid model already provides a strong fit for the empirical data when β is 0.594. However, adding a second parameter, ϵ , gives the hybrid-with-error model additional flexibility to slightly adjust the predictions made by particular instantiations of the simpler model. It is highly likely that the interactions that result from adjusting both parameters at once will yield at least some model instantiations where the fit improves in more conditions than it worsens.

Impressionistically, this is because the model has two strategies available for adjusting predictions. Consider the unmodified conditions. Given a conceptual “starting point” with $\beta = 1$ and $\epsilon = 0$, the predicted probability of concluding the polar interpretation is 1 in all three conditions. The model could raise ϵ to move all three interpretations closer to 0.5, or it could lower β to move the predicted interpretations closer to the empirical prior distribution

for the three conditions. It is most likely that the best result involves some combination of raising ϵ and lowering β , so it would not be surprising if the joint likelihood is more optimally maximized by the hybrid model with error than the hybrid model without error, which only has the option of adjusting β . This prediction for the maximized joint likelihood is borne out.

So, an ideal model comparison technique will penalize models for containing additional parameters. In essence, the goal is to select the simplest possible model that provides a reasonable fit for the empirical data. While the best fit for a particular data set is likely to come from a model with a high number of free parameters, it is relatively more likely that complex models provide high likelihoods in individual instantiations by chance, but actually do not model the real-world mechanisms underlying the generation of the data. The goal is to identify a model whose parameters can be justified by the goodness of fit they provide across a wide range of possible parameter values, as opposed to a model with very strong fit in a narrow range of possible parameter values, but poor fit otherwise.

One measure that is known to strike a balance between prioritizing good fit and penalizing unnecessary parameters is the *Bayes factor* (Jeffreys, 1961; Kass and Raftery, 1995; Qing and Franke, 2015). In its simplest form, the Bayes factor is simply a ratio of a particular measure taken for two models called the *marginal likelihood* or *model evidence*. Equation 4.11 gives the definition of the marginal likelihood, ML, for a particular model $\mathcal{M}$.

$$\text{ML}_{\mathcal{M}} = \int P(\theta) \cdot P(D|\mathcal{M}, \theta) \text{d}\theta \quad (4.11)$$

In this equation, θ represents a particular set of values for each of the free parameters. For a model with just a parameter ϵ , θ could be 0, 0.01, 0.05, and so on. For a model with multiple parameters, it would represent a tuple of parameter values – for example, (β, ϵ) pairs such as (0,0), (1,0), (0.5,0.05), and so on.

$P(D|\mathcal{M}, \theta)$ represents the probability of generating a particular set of empirical data, D , given the model $\mathcal{M}$ and the set of parameter values specified as θ – in other words, the

likelihood of that particular instantiation of the model. Note that D and $\mathcal{M}$ here abstract over the conditions of the experiment; D is the full set of experimental data, and $\mathcal{M}$ is the model for generating the predicted data across all conditions. $P(\theta)$ represents the probability of that particular set of parameter values being the case according to some prior probability distribution for each parameter.

Intuitively, the marginal likelihood represents a weighted average of model likelihoods across the entire possibility space for parameter values.⁵ The average is weighted in that the likelihood of each model instantiation is multiplied by the probability of the specific parameter values holding according to their prior probability distributions. Importantly, the marginal likelihood will be low if a particular model provides poor fit across much of the parameter space, and higher when the fit is good across a large portion of the space. Thus, the marginal likelihood can give some insight into which models provide a strong fit in a particular instantiation by luck or overfitting, and which are more consistently capable of modeling the empirical situation.

To use Equation 4.11 to calculate the marginal likelihood of a particular model, two issues must be addressed: the prior probability distribution for the parameter space itself (referred to as a *hyperprior*) and the limits of integration.

The factor $P(\theta)$ reflects the assumption that certain combinations of parameter values are *a priori* expected to be more likely, allowing the marginal likelihood calculation to weight the likelihood of a particular model instantiation by how likely it was considered according to prior probabilities. It is not clear that such *a priori* expectations should exist for the current β parameter. To the extent that the hybrid model is the correct model of VPE interpretation, there have been few firm proposals in the literature on how information from

5. To understand this, consider that any weighted average can be calculated as $\frac{\int_a^b w(x) \cdot f(x) dx}{\int_a^b w(x) dx}$, for function f and weights $w(x)$. In this case, the function values of interest are the model likelihoods, and the weight is $P(\theta)$. By definition of a probability density function, $\int P(\theta) d\theta$ across the entire range of possible values of θ must be equal to 1, so the denominator of the weighted average expression is eliminated when the weights are determined by a probability distribution and the limits of integration span the full range of possible outcomes.

an identity strategy and from a discourse strategy are combined to arrive at an interpretation of a VPE site. It is not clear that the parameter value should be considered more likely to bias identity-driven interpretation or discourse-driven interpretation. For this reason, the calculations below use a “flat” prior for β , meaning each of the possible values are treated as equally likely (Qing and Franke, 2015).

For ϵ , it seems intuitive that linguists would have some *a priori* expectations about the probability of particular values. It seems likely that the proportion of times listeners guess at an interpretation completely at random during utterance interpretation is quite low, even in a potentially difficult interpretation task like the one completed in Experiment 7. It would be reasonable to propose that listeners might guess 1%, 2%, or even 5% of the time, but not that they would guess, say, 75% of the time. This seems to support a prior distribution over ϵ with the bulk of the probability near 0.

The issue of selecting a prior distribution for ϵ intersects with the question of how to set the limits of integration. For reasons that will be discussed immediately below, the current analysis will restrict ϵ to take values between 0 and 0.2, even though the theoretical range of ϵ values ranges from 0 to 1. Given this restriction, it is somewhat less clear what the *a priori* expectations for the probability of different values of ϵ should be. Perhaps the bulk of the distribution should still be close to 0, but “high” values within the interval 0 to 0.2 are certainly more reasonable than “high” values within the interval 0 to 1.

For the analysis presented below, a flat prior for ϵ will be used, representing equal probability for any value of ϵ in the interval 0 to 0.2. However, it is important to note that selecting a different prior distribution over ϵ could have substantial effects on the marginal likelihoods for models containing ϵ . Future work could focus more on considering models with different distributions, particularly in light of empirical evidence suggesting how likely it is that listeners engage in random behavior in utterance interpretation.

The limits of integration for the marginal likelihood calculation are simply the lower and upper bounds of any free parameters represented in the equation. For the current β

parameter, this is straightforward; since β represents a proportion, the values considered will range from 0 to 1, inclusive. (Recall that the β values considered in the maximum likelihood estimation ranged from 0 to 0.99. This is because when β is 1, the likelihood in the unmodified conditions is 0, meaning the log likelihood was $-\infty$, making it impossible to make meaningful comparisons using the sum of the log likelihoods in each condition. Since the marginal likelihood calculation considers likelihood rather than log likelihood, β can be 1.)

Determining the limits of integration for ϵ is more complicated, as alluded to above. In theoretical terms, ϵ should likewise be allowed to range from 0 to 1, since it represents a proportion. However, analysis using these limits leads to a conclusion that is not supported by domain-specific knowledge. Recall that the identity model with error has maximized likelihood with an ϵ of about 0.3. By contrast, the optimal ϵ value for other models containing this parameter is closer to 0.1. Since the marginal likelihood averages model likelihoods across the entire parameter space, the identity model with error systematically outperforms the other models for medium and high values of ϵ .

In effect, when the limits of integration for ϵ are set to 0 and 1, the identity model with error actually has the highest marginal likelihood. However, this is a result of considering a wide range of ϵ values that are likely to be judged unreasonable by linguists. Since ϵ represents the proportion of time listeners interpret utterances at random, it seems unlikely that its real-world value should be a high proportion, like 0.7, or even a “medium” proportion like 0.5 or 0.25. It is exactly these values of ϵ that drive this model’s advantage in marginal likelihood, while it misses the empirical data quite badly for more reasonable values of ϵ .

For this reason, the limits of integration for ϵ were set to 0 and 0.2, thus considering a maximum percentage of random guessing of 20%. It is important to note that this was done with the express intention of limiting the marginal likelihood for the identity model with error. However, this limitation is theoretically motivated by domain-specific knowledge – it is simply not plausible to propose that listeners guess an interpretation at random such

a large proportion of the time. As for the prior distribution over ϵ , future research could further explore the model predictions for different ranges of possible ϵ values to determine how the model comparison changes for different theoretically motivated estimates of random behavior.

With the prior distributions and limits of integration for the current free parameters set, the equations giving the marginal likelihoods for each model can be set. First, consider the models with no free parameters, the identity model without error and the salience model without error. Since there are no free parameters, the joint likelihood for these models is constant, and integration is not necessary. The equation for marginal likelihood is shown in Equation 4.12.

$$\text{ML}_{\mathcal{M}} = P(D|\mathcal{M}) \quad (4.12)$$

Next, consider the models with only the free parameter ϵ . The limits of integration were set to 0 and 0.2. Thus, the marginal likelihood for a model with only an ϵ parameter is given by Equation 4.13.

$$\text{ML}_{\mathcal{M}} = \int_0^{0.2} P(\epsilon) \cdot P(D|\mathcal{M}, \epsilon) d\epsilon \quad (4.13)$$

For a model with only a β parameter, the limits of integration are 0 and 1, and the marginal likelihood is given by Equation 4.14.

$$\text{ML}_{\mathcal{M}} = \int_0^1 P(\beta) \cdot P(D|\mathcal{M}, \beta) d\beta \quad (4.14)$$

Finally, for the model with two parameters, the marginal likelihood is given by Equation 4.15.

$$\text{ML}_{\mathcal{M}} = \int_0^{0.2} \int_0^1 P(\beta, \epsilon) \cdot P(D|\mathcal{M}, \beta, \epsilon) d\beta d\epsilon \quad (4.15)$$

Recall that the marginal likelihood is an average of the likelihoods for all different instantiations of the model, weighted by the probability of the combination of priors determining the given instantiation. In this case, flat priors are being used for both β and ϵ , meaning for each model, every instantiation is equally likely. This means that the marginal likelihood formulas shown in Equations 4.13 through 4.15 are equal to the simple arithmetic mean of the likelihoods of all individual instantiations of the respective model. (This would not be the case if any different prior were used for one of the parameters, so it is worth considering what effect changing the shape of these prior distributions would have on the model comparison discussed below.)

Here, the marginal likelihood for each model in each condition was approximated by using the model formula and the empirical priors to calculate the likelihood for many different model instantiations. For the models with no free parameters (identity without error and salience without error), no approximation was necessary; the likelihood was directly calculated using the formula. For models with only an ϵ parameter, the likelihood was calculated for 2001 instantiations of the model with ϵ ranging from 0 to 0.2 in equal steps; the arithmetic mean of the 2001 likelihoods was then taken. Similarly, for the hybrid model without error, the marginal likelihood was calculated as the arithmetic mean of 2001 instantiations of the model with β varying in equal steps from 0 to 1. For the hybrid model with error, the marginal likelihood was estimated as the arithmetic mean of 40,401 model instantiations. The instantiations considered were determined using a “grid” of parameter values with ϵ varying from 0 to 0.2 in 201 equal increments, and β varying from 0 to 1 in 201 equal increments.

The nine experimental conditions were considered separately for each model in this analysis. While the likelihood of a particular instantiation of a model generating the empirical data across the entire experiment can be calculated by multiplying the likelihood in each of the nine conditions together, this product can be misleading in some circumstances. In particular, recall that the identity model without error has 0 likelihood of generating the

experimental posterior data with an unmodified antecedent, since it predicts a probability of 0 of selecting the scalar interpretation, whereas this interpretation was selected in the experiment. Multiplying this model’s likelihoods in the other conditions by its likelihoods in the unmodified conditions results in a joint likelihood of 0. This is correct on a theoretical level, but it obscures the fact that the model performs fairly well in the other 6 conditions. Beyond this, analyzing the conditions separately helps localize the strengths and weaknesses of the various models in a theoretical context.

With the marginal likelihoods for each model in each experimental condition known, the Bayes factors can be calculated. Recall that the Bayes factor is simply the ratio of the marginal likelihoods of two models, as shown in Equation 4.16.

$$\text{BF}_{\mathcal{M}1\mathcal{M}2} = \frac{\text{ML}_{\mathcal{M}1}}{\text{ML}_{\mathcal{M}2}} \quad (4.16)$$

In this case, the marginal likelihood for Model 1 in a particular condition will be compared to the marginal likelihood for Model 2 in the same condition. Thus, comparing two models will consist of calculating nine Bayes factors, one for each experimental condition. Table 4.3 shows some guidelines for interpreting the Bayes factor comparing two models, based on categories $\frac{1}{2}$ a unit wide on the scale of $\log_{10}(\text{BF})$ (adapted from Jeffreys, 1961; Kass and Raftery, 1995). Note that since the Bayes factor is a ratio, a value of 1 indicates perfect indifference between the models, values over 1 indicate stronger evidence for Model 1 (the numerator model), and values under 1 indicate stronger evidence for Model 2 (the denominator model).

Here, the analysis focuses primarily on comparing the two hybrid models to the identity and salience models. Bayes factors showing comparisons between the non-hybrid models are shown in Appendix H. Table 4.4 shows the Bayes factors by condition comparing the hybrid model without error (β only) to four models: identity without error, identity with error, salience without error, and salience with error. In each case, the hybrid model without error is Model 1 (the numerator), so Bayes factors over 1 indicate evidence in favor of the hybrid

$BF_{\mathcal{M}1\mathcal{M}2}$	Interpretation
<0.01	Decisive evidence in favor of $\mathcal{M}2$
0.01 to 0.1	Strong evidence in favor of $\mathcal{M}2$
0.1 to 0.316	Substantial evidence in favor of $\mathcal{M}2$
0.316 to 3.16	Not worth more than a bare mention
3.16 to 10	Substantial evidence in favor of $\mathcal{M}1$
10 to 100	Strong evidence in favor of $\mathcal{M}1$
>100	Decisive evidence in favor of $\mathcal{M}1$

Table 4.3: Bayes factor interpretation guidelines.

<i>Antecedent</i>	<i>Context</i>	<i>Model 2</i>			
		Identity, no ϵ	Identity + ϵ	Saliency, no ϵ	Saliency + ϵ
Exophoric	Unavailable	1.00	1.39	1.00	1.39
Exophoric	Available	1.00	0.583	1.00	0.583
Exophoric	Salient	1.00	2.04	1.00	2.04
Unmodified	Unavailable	∞	45.0	38.1	125
Unmodified	Available	∞	2.45e04	5.62	12.9
Unmodified	Salient	∞	1.85e18	6.84e08	4.80e06
Modified	Unavailable	1.00	0.439	1.00	0.439
Modified	Available	1.00	0.766	1.00	0.766
Modified	Salient	1.00	1.17	1.00	1.17

Table 4.4: Bayes factors by condition with hybrid model without error as Model 1.

model over the identity or saliency models.

There are some special circumstances underlying some of these Bayes factors that are worth reviewing. First, recall that the identity model without error has a likelihood of 0 of generating the empirical data in the unmodified-antecedent conditions. Since the hybrid model without error has non-zero likelihood in these conditions, the Bayes factor comparing these models in these conditions has denominator 0 and is shown in the table as ∞ .

Second, the identity, saliency, and hybrid interpretation strategies make theoretically identical predictions in the exophoric and modified-antecedent conditions – that is, the compatibility rating of each interpretation with the VPE reply is 1, and interpretation is driven by the prior availability of each interpretation. Since the identity model without error and the saliency model without error have no free parameters, and the hybrid model without error only combines the predictions of these models in some proportion, the marginal like-

lihoods of each model are identical in the exophoric and modified conditions are identical, leading to a Bayes factor of exactly 1.

Similar circumstances underlie the identical Bayes factors comparing the hybrid model to the identity-with-error model and the hybrid model to the salience-with-error model in these conditions; the two models have identical likelihoods for corresponding values of ϵ , so the Bayes factors comparing the hybrid model to each in these conditions are identical.

Beyond these observations, the Bayes factors provide compelling evidence in favor of the hybrid model (without error). In the exophoric and modified conditions, the comparison is generally indifferent between the hybrid model and any other model, with all values between 0.316 and 3.16, the “not worth more than a bare mention” range. However, in the unmodified-antecedent conditions, the evidence in favor of the hybrid model ranges from “substantial” (versus the salience model without error, unmodified/available condition) to “strong” (versus identity with error and salience without error in the unmodified/unavailable condition and versus salience with error in the unmodified/available condition) to “decisive” in all other unmodified-antecedent comparisons.

Table 4.5 likewise shows the Bayes factors by condition comparing the hybrid model with error to each of the identity and salience models, with the hybrid model as Model 1, the numerator model. Note that the Bayes factors reported here as 1.00 are rounded; the model predictions between the hybrid model with error and the identity-with-error and salience-with-error model are no longer identical as they were for the hybrid-without-error model, but are very close.

The Bayes factors comparing the hybrid model with error to the identity and salience models are qualitatively identical to those for the hybrid model without error. In the exophoric and modified-antecedent conditions, the comparison reflects indifference between the models. In the unmodified conditions, however, the Bayes factors indicate substantial, strong, or decisive evidence in favor of the hybrid model, depending on the condition and the model being compared.

<i>Antecedent</i>	<i>Context</i>	<i>Model 2</i>			
		Identity w/o ϵ	Identity w/ ϵ	Saliency w/o ϵ	Saliency w/ ϵ
Exophoric	Unavailable	0.717	1.00	0.717	1.00
Exophoric	Available	1.72	1.00	1.72	1.00
Exophoric	Salient	0.490	1.00	0.490	1.00
Unmodified	Unavailable	∞	45.8	38.7	127
Unmodified	Available	∞	2.53e04	5.81	13.3
Unmodified	Salient	∞	2.00e18	7.41e08	5.19e06
Modified	Unavailable	2.28	1.00	2.28	1.00
Modified	Available	1.30	0.999	1.30	0.999
Modified	Salient	0.853	0.999	0.853	0.999

Table 4.5: Bayes factors by condition with hybrid model with error as Model 1.

<i>Antecedent</i>	<i>Context</i>			
		Unavailable	Available	Salient
	Exophoric	0.717	1.72	0.490
	Unmodified	1.02	1.03	1.08
	Modified	2.28	1.30	0.853

Table 4.6: Bayes factors by condition with hybrid model with error as Model 1 and hybrid model without error as Model 2.

Finally, Table 4.6 shows the Bayes factors comparing the two hybrid models directly in each condition, with the hybrid-with-error model as Model 1.

These values point to indifference between the exclusion or inclusion of the error parameter, ϵ , in the hybrid model in all conditions. In some conditions, such as modified/unavailable, the model with error performs somewhat better, while in others, like the exophoric/salient condition, the model without error is better, but none of the Bayes factors even attain “substantial” evidence for one model or the other.

4.3.9 Discussion

The model comparison findings are broadly in line with the conclusions drawn above from visual inspection of the plots of the maximum likelihood instantiation of each model against the empirical posterior data. In the exophoric and modified conditions, the predictions of each model have the same underlying architecture, and each model generally provides a good

fit to the data. However, in the unmodified conditions, both the identity models and the salience models miss the empirical data badly, especially in the unmodified/salient condition. The Bayes factors reflect the intuition that the hybrid model provides a substantially better fit in these conditions, and the advantage is especially strong in the unmodified/salient condition.

Considered together, the results point to a definitive advantage of the hybrid models over any of the identity-only or salience-only models. There is substantially stronger evidence for the hybrid models in the conditions with an unmodified antecedent. Although the hybrid models were not always stronger in the exophoric and modified conditions, no disadvantage compared to an identity-only or salience-only model attained the Bayes factor threshold of “substantial” evidence. Furthermore, since the calculation of marginal likelihood implicitly penalizes models that include extra parameters to provide a good fit in a few instantiations, but that largely miss the empirical data otherwise, the success of the hybrid models cannot readily be attributed to the inclusion of unnecessary parameters relative to the simpler models. Based on the strong evidence in favor of the hybrid models in the unmodified conditions, and taking into account that they perform roughly comparably to the identity-only and salience-only models in the other conditions, it seems safe to conclude that the hybrid models are substantially better models of the interpretation behavior in Experiment 7.

By contrast, direct comparison of the hybrid model without error and the hybrid model with error was inconclusive. There was relatively stronger evidence for the model without error in some conditions, and stronger evidence for the model with error in others. According to the interpretation guidelines in Table 4.3, the Bayes factor did not indicate a meaningful difference in the evidence for either model in any condition.

Since the model with error contains an additional parameter compared to the model without error, the marginal likelihood calculation implicitly penalizes it for instantiations in which the ϵ value leads to a poor fit. In this case, the Bayes factors indicate that the overall evidence was not highly affected by the inclusion of this parameter. However, a general

preference for the simplest possible model would suggest that the error term is not necessary to provide a suitable model of interpretation behavior in Experiment 7. Here, it is worth recalling the earlier decisions that were made about the analysis of the ϵ parameter, notably the decision to limit its range from 0 to 0.2 and the use of a flat hyperprior. Altering either of these analytical choices could affect the marginal likelihoods for the hybrid model with error and could potentially lead to a substantial evidence advantage or disadvantage relative to the hybrid model without error.

4.4 General discussion

The experiments presented in the previous chapter provided useful information about the mechanisms underlying interpretation, constraining the hypothesis space at both the identity and discourse ends of the interpretive continuum. However, experimental methodologies can be limited in their utility, especially in cases where the predictions of two hypotheses are difficult to tease apart. This is the case for the “permissive” identity account and the “complex” discourse account, which both allow for interpretation patterns that are intermediate between the predictions of the strict identity and simple discourse accounts.

Beyond providing additional evidence for the conclusions drawn in the previous chapter, the modeling work presented here outlines another methodology that can be used to probe mechanisms of context-situated VPE interpretation. The hybrid models as presented here do not distinguish between the permissive identity and complex discourse accounts, but the modeling methodology as a whole provides a frame that may lead to future modeling work that can adjudicate between the two.

To summarize, the chapter opened with Experiments 5 through 7. These experiments collected forced-choice judgments of the availability of number-unmodified and number-modified propositions in the prior discourse and as ellipsis interpretations in the experimental paradigm from the previous chapter. The main goal was to collect proportion measurements that would be used as inputs to the probabilistic models in the case of the priors, or as an

empirical standard of comparison for the model outputs in the case of the VPE interpretations. Comparison of the results of the prior and posterior experiments replicated the core interpretation from the previous chapter. VPE interpretation is affected by information in the broader discourse when a linguistic antecedent is present, but not to the extent that the effect of the antecedent is reduced to its mere contribution to salience.

The experiments in this chapter also yielded potentially useful insights about the utility of Likert score rating tasks versus forced-choice tasks. When one interpretation was significantly preferred over the other, the forced-choice task represented fine gradations in preference in cases where they were obscured in the Likert task, such as the modified-antecedent conditions for VPE interpretation. Similarly, the results of Experiment 7 showed suggestive numeric trends of increasing consideration of number information when switching from an unavailable to an available context, even when there was a linguistic antecedent. This trend was not observed in Experiment 1, and while it did not attain significance in Experiment 7, it suggests that there may be a subtle effect of exclusively nonlinguistic manipulations on VPE interpretation with viable antecedents that was not detected here. Future research may be able to refine these experimental paradigms to detect effects of exclusively nonlinguistic contexts, eliminating the confound of the linguistic information that was visually represented in the salient comic strips in the current setup.

With these three experiments completed, attention turned to the modeling technique itself. Six models were constructed with different specifications for transforming the prior proportions measured in the experiments to posterior probability distributions for accessing a number-unmodified or -modified interpretation for ellipsis. The models each schematized one of three interpretive strategies: identity, considering only the linguistic antecedent; salience, considering only the prior discourse salience of number-unmodified versus -modified propositions; or hybrid, combining the predictions of the former two in proportion. For each strategy, a model with and without an allowance for random guessing – whether in naturalistic interpretation broadly construed or the narrow experimental task – were considered.

Two comparison techniques were used to adjudicate between the models. First, maximum likelihood estimation was used to determine the parameter values for each model that were most likely to generate the Experiment 7 data, allowing impressionistic examination of the “peak” fit for each model. Second, the Bayes factor was calculated to compare both hybrid models to the remaining models. This abstracts over all possible parameter values, implicitly penalizing models that overfit by including extraneous parameters that fit the data in a small number of instantiations at the expense of poor fit elsewhere. Both techniques indicated that the hybrid models provided a better fit to the empirical data than any of the identity or salience models. Unsurprisingly, this was largely driven by the predictions and results in the unmodified-antecedent conditions, and especially the unmodified/salient condition that drove the critical conclusions of the previous chapter.

The modeling results were interpreted as supporting the main conclusions of the previous chapter: the strict identity account, represented by the identity models, and the simple discourse account, represented by the salience models, are inadequate accounts of VPE interpretation in context. The strength of the hybrid model indicates that the correct model of VPE interpretation must “combine” the predictions of the two simpler accounts – that is, the strict identity account must be made sensitive to discourse, and the simple discourse account must be constrained by the form of the antecedent beyond its contribution to salience.

However, the success of the hybrid model does not lead to a clear conclusion about the status of broader discourse information in the grammar of VPE. In principle, the hybrid models as constructed here could represent either the permissive identity account or the complex discourse account. In the case of an identity account, the β parameter would schematize the relative consideration that an alternative, discourse-derived antecedent is given relative to the overt antecedent. A large value of β would indicate that the overt antecedent is given more consideration, while a smaller value lends more weight to the accommodated antecedent. Likewise, for the complex discourse account, β simply controls the weight that processing heuristics give to antecedent information rather than to information from the

discourse at large.

It is worth noting that the models provided here are only a sketch of the possible interpretive mechanism for VPE. Models dealing with more articulated interpretive mechanisms and their predictions could alter the specification of existing components or add different components entirely. For instance, when testing the complex discourse model, the “discourse at large” component could be replaced with a “non-antecedent” component. In the current paradigm, the contribution of the linguistic antecedent is included in the prior probabilities as a measure of the salience of competing interpretations in the context at large. If salience instead were measured when only the nonlinguistic context and prior linguistic context were available, but not the linguistic antecedent, the model specification and parameter values would reflect a different division of labor, between the antecedent itself and all of the information available outside the antecedent. Similarly, additional mathematical components could be included in the model to represent other interpretation heuristics, such as Hardt’s (1993) heuristics for identifying a preferred linguistic antecedent within a complex linguistic discourse.

A related issue that is worth mentioning here is the relationship between the β parameter and real-world strategies for interpretation. Recall that the hybrid model without error, without simplifying the salience component, appears as in Equation 4.17.

$$P(m|u, c) = \beta \cdot \frac{\llbracket u \rrbracket_{\text{identity}}^{m,c} \cdot P(m|c)}{\sum_{m' \in M} \llbracket u \rrbracket_{\text{identity}}^{m',c} \cdot P(m'|c)} + (1 - \beta) \cdot \frac{\llbracket u \rrbracket_{\text{salience}}^{m,c} \cdot P(m|c)}{\sum_{m' \in M} \llbracket u \rrbracket_{\text{salience}}^{m',c} \cdot P(m'|c)} \quad (4.17)$$

This model schematizes interpretation as a proportional combination of the predictions of the identity and salience strategies, with a proportion of β of the posterior probability distribution determined by the identity strategy and a proportion of $1 - \beta$ determined by the salience strategy. However, there are actually two different behaviors that this could describe. The first possibility is that listeners randomly select between faithfully executing

one of the two simpler interpretation strategies, with a probability of β that they use the identity strategy. The second is that listeners actually calculate the predictions of both strategies and combine them in proportion.

If this second possibility is correct, then the model shown in Equation 4.17 is not strictly valid. This model uses the denominators to normalize the viable interpretations into proper probability distributions separately from one another before combining them with the β parameter. This is valid for the random choice strategy, but not for the proportional combination strategy. In the latter, the core calculation (numerator) should include combining both strategies before normalizing them with respect to the strength of all remaining interpretations according to this strategy. The corresponding model is shown in Equation 4.18.

$$P(m|u, c) = \frac{\beta(\llbracket u \rrbracket_{\text{identity}}^{m,c} \cdot P(m|c)) + (1 - \beta)(\llbracket u \rrbracket_{\text{salience}}^{m,c} \cdot P(m|c))}{\sum_{m' \in M} \beta(\llbracket u \rrbracket_{\text{identity}}^{m',c} \cdot P(m'|c)) + (1 - \beta)(\llbracket u \rrbracket_{\text{salience}}^{m',c} \cdot P(m'|c))} \quad (4.18)$$

Quantitatively the behavior of this model is likely to be very similar to that of the hybrid models considered above. However, this serves as an illustrated example of the nuanced interaction between model specifications and real-world behavior on the part of listeners. It will no doubt be a challenging endeavor to harness the power of the modeling methodology outlined here to study more complex accounts of interpretation, such as models that consider nonlinguistic information as well as more complex linguistic discourses than were included in the experiments here.

In closing, the modeling investigation conducted in this chapter provided evidence from a second domain in favor of the conclusion from the last chapter that a permissive identity model and a complex discourse model, but not a strict identity account or a simple discourse account, remain viable models of VPE interpretation in context. While the models as constructed here could not adjudicate between the two viable accounts, they provide the building blocks for additional modeling investigation of more nuanced predictions made by

each.

This type of methodology might be especially important for testing future developments of a complex discourse account of VPE interpretation. The predictions of such an account can be complex, and may be difficult to assess directly with an experimental paradigm. Incorporating additional components representing proposed interpretation heuristics into the probabilistic model can provide an additional testing ground for validating or refining these predictions. Beyond this, an additional future direction might be to model the interpretation of other verbal anaphors, like *do that*, using similar models. Investigation of how the same model architecture handles the empirical facts of interpretation for different anaphors, with consideration of the parameter values that result for each type of construction, may lead to valuable insights about the nature of anaphora and context dependence more generally.

CHAPTER 5

BACKGROUND: GIVENNESS, INFERABILITY, AND DEACCENTING

The second half of this dissertation focuses on a phenomenon that will be referred to here as *anaphoric deaccenting*. Similarly to verb phrase ellipsis, this phenomenon raises fundamental questions about the nature of the connections between context-sensitive expressions, their linguistic antecedents, and broader representations of discourse that are the focus of this dissertation. In the case of deaccenting, the main question that will be addressed is whether deaccenting is licensed by broad discourse considerations – accessibility, givenness, inferability – or by identity with a linguistic antecedent supplemented by an independent pragmatic mechanism deriving more peripheral cases.

The basic concept of *deaccenting* assumes that languages specify a “default” pattern of accentuation for sentences uttered in a neutral discourse context (e.g., Chomsky and Halle’s (1968) Nuclear Stress Rule). In English, the most reliable “default” pattern is that the most embedded contentful constituent in a clause will receive a *nuclear pitch accent*, meaning it is prosodically emphasized to the point of being perceived as the most prominent constituent in the clause (Selkirk, 1984; Büring, 2016).

Constituents that are in a position to receive a pitch accent under default conditions are said to be *deaccented* when they are realized without that pitch accent. This happens most often when the constituent is identical to material that has been pronounced in the recent linguistic context. Thus, deaccenting is *anaphoric* in the sense that the decision of whether to deaccent a particular constituent is a function of other linguistic and possibly nonlinguistic information that is available in the discourse.

Examples (59) and (60) demonstrate a contrast between an accented constituent and a deaccented constituent. In this dissertation, small capital letters will denote that a material is accented – associated with phrasal prosodic prominence – while italics will denote that

a constituent is deaccented – produced with a low or absent pitch accent. This annotation will be selectively applied to critical clauses used to illustrate prosodic realizations.

(59) The caterer asked if there were any drinks to avoid, and I said I don’t like COFFEE.

(60) The caterer asked if they should supply coffee, and I said I don’t LIKE *coffee*.

In a canonical realization of (59), COFFEE is realized with prosodic prominence because, intuitively, it answers the question of which drinks to avoid. COFFEE is also in a position to receive a nuclear pitch accent in a “neutral” discourse if the final clause were uttered in isolation. By contrast, in (60), *coffee* is canonically deaccented because it is identical to the instance of *coffee* in the antecedent clause. In (60), the nuclear pitch accent has been “shifted” from default position to the preceding contentful constituent, LIKE, whence another name for deaccenting that has been used in the literature, *stress shift*.

It has long been recognized that there is a reliable mapping between the discourse-accessibility of a constituent and its prosodic realization, with new or “contrastive” material generally being associated with prominence, and antecedent-identical material being de-emphasized (Chafe, 1974; Pierrehumbert and Hirschberg, 1990, i.a.). However, there are more possible levels of discourse-accessibility than simply “given” and “new.” Of particular interest to this dissertation is “accessible” or “inferable” material – that is, constituents whose meanings that are entailed by the content of an antecedent, can reasonably be accommodated on the basis of the antecedent, are made available by the nonlinguistic context, and so on (Prince, 1981; Baumann and Riester, 2012).

It is commonly claimed in literature narrowly focusing on the prosody of accent, as well as literature focusing on the grammar of deaccenting, that such accessible or inferable constituents can be deaccented (Lakoff, 1968; Chafe, 1974; Rooth, 1992a; Tancredi, 1992; Selkirk, 1995; Fox, 2000; Baumann and Riester, 2012; Büring, 2016; Chodroff and Cole, 2019, i.a.). Several different types of accessible material have been proposed to be deaccentable. First, constituents which co-refer with an antecedent, despite receiving a different linguistic encoding, can be deaccented, as shown in (61).

(61) A: Did you see Dr. Cremer to get your root canal?

B: Don't remind me. I'd like to STRANGLE *the butcher*.

(Büring, 2007)

Here, *the butcher* and *Dr. Cremer* refer to the same entity, and *the butcher* is treated as given information and deaccented despite being formally new to the discourse.

Constituents are also said to be deaccentable when, modulo type shifting via existential closure, they are entailed by a constituent from the antecedent, as in (62).

(62) Bach wrote many pieces for viola. He must have LOVED *string instruments*.

(van Deemter, 1999)

In this case, *string instruments* is a superset of *viola*, so the prior mention of *viola*, intuitively, entails that the concept of a string instrument is available in the discourse.

Interestingly, there are extended examples similar to (61) type that do not require a formal entailment relation. Constituents are also noted to be deaccentable which are agreed by the discourse participants to be reasonably concluded from an antecedent, even if they are not strictly entailed. Often, these examples require fairly sophisticated world knowledge to support the inference that the meaning of the deaccenting target is accessible. (63) is an example.

(63) First John called Mary a Republican, and then SHE *insulted* HIM.

(Lakoff, 1968)

Here, deaccenting *insulted* is argued to be felicitous when both discourse participants agree to an implicit assumption that calling someone a Republican constitutes insulting them. If this assumption holds, then the meaning of *insult* can be treated as given in the discourse.

Finally, although such cases have not received much attention, it has been argued that deaccenting is possible when supported by salient information from the nonlinguistic context, as in (64).

(64) [Hearer cocks their head to one side as if listening for a faint or distant noise.]

Speaker: I *heard it, too*.

(Rochemont, 1986)

Each of (61) through (64) constitutes deaccenting under *nonidentity* with the prior linguistic context, in contrast to (60), where the deaccented constituent has an identical correlate in the prior linguistic context. While judgments of this type of example as acceptable are more or less uniform across the literature, there has been surprisingly little rigorous empirical investigation of the deaccenting of discourse-accessible material. Given the pervasiveness of these examples, it would seem to be worthwhile to verify that such deaccenting occurs in production, and that naive listeners judge it to be felicitous in perception.

One of the first studies to systematically investigate the prosody of discourse-accessible material was a production study conducted by Chodroff and Cole (2019), roughly concurrently with the work presented in this dissertation. Chodroff and Cole manipulated the antecedents in their production study so that a target noun could have one of four levels of discourse status: given (*marmalade-marmalade*), accessible (*preserves-marmalade*), new (*make everything from scratch-marmalade*), or contrastive (*jam-marmalade*).

Chodroff and Cole's results were problematic given the judgments reported in the prior literature. In trained annotation of pitch accent, accessible nouns received a high or rising pitch accent significantly less often than new or contrastive nouns, but also significantly more often than given nouns. In a phonetic analysis, accessible nouns were produced reliably differently from given nouns in both duration and intensity, and were significantly less likely to be realized entirely with creaky voice. By contrast, the phonetic differences between accessible and new or contrastive nouns were less reliable, with contrastive nouns realized with higher intensity than accessible ones, and new nouns more likely to have fully modal voicing rather than a mix of modal and creaky voicing.

These results largely conflict with the judgments reported in the literature, where deaccenting of accessible or inferable material is claimed to be fully acceptable and deaccenting

is often modeled as being mandatory when it is possible. Chodroff and Cole's results show a clear separation between the behavior of accessible and given nouns, where given nouns were less emphasized than accessible nouns across all of their phonological and phonetic variables. By contrast, the distinction between new and contrastive nouns, which canonically are expected to be accented, and accessible nouns was less clear, with a quantitative difference in the phonological measure and only sporadic differences in the phonetic analysis.

This finding suggests that deaccenting of accessible material is far from mandatory and motivates additional empirical investigation on the prosodic behavior of inferable constituents. However, the problem is more than an empirical one. Researchers of the interfaces between prosody and syntax, semantics, and discourse have devoted considerable attention to the development of a grammatical constraint characterizing which constituents can (or must) be deaccented. Of interest to this dissertation is the treatment of nonidentical material – that is, deaccented expressions that do not have an identical correlate in the linguistic antecedent.

This chapter reviews two main strategies for accounting for the relationship between constituents' givenness, accessibility, and novelty and their prosodic realization. The first strategy is referred to here as the *grammatical* approach to deaccenting under nonidentity. Accounts of this type make direct reference to inferability or givenness in the grammar of deaccenting. That is, it is precisely the fact that a constituent is given or inferable in the discourse, and not identity with a linguistic antecedent, that licenses its deaccentuation.

Theories in the grammatical camp take a wide variety of approaches to characterizing givenness and, by extension, deaccentable material. One approach seeks to provide a unified account of both givenness and the broader, independently necessary theory of focus. Focus and the assignment (or lack thereof) of pitch accent within a sentence are inextricably linked to one another. Thus, some researchers aim to give a focus-theoretic analysis of givenness that can characterize deaccentable material using independently motivated theoretical machinery. Work in this vein often provides a structural definition of givenness that is calculated

with respect to a previous linguistic antecedent.

A second line of work taking the grammatical approach to deaccenting under nonidentity focuses on more holistic, discourse-driven assessments of givenness. These accounts may eschew more structural definitions of givenness, concentrating instead on givenness marking as a strategic communicative decision that does not have formal licensing conditions. On this view, deaccentable material is exactly that which the speaker wishes to index as given in the discourse, regardless of whether it has a formal antecedent. Often, givenness marking is still assumed to interact with focus theory at another level to determine the full prosodic realization of a sentence, but the details of this interface are not always fully spelled out.

The second broad approach to deaccenting strategy will be referred to here as the *accommodation* approach to deaccenting under nonidentity. This thread is somewhat less represented in the literature, and holds that deaccenting is only grammatical when the target has an identical correlate in the antecedent. This does not mean, however, that deaccenting under nonidentity is unavailable, or even infelicitous in most cases. Rather, this account proposes that such cases are generated via an extragrammatical process that constructs an alternative antecedent containing a correlate identical to the target. In essence, from a listener's perspective, hearing deaccenting that is strictly ungrammatical serves as a trigger to accommodate a new structure that, had it been the actual antecedent, would have licensed the deaccenting.

The grammatical and accommodation approaches to the deaccenting of antecedent-nonidentical material constitute fundamentally different characterizations of the connection between a deaccenting target and the broader discourse. According to the grammatical approach, the grammar of deaccenting licensing makes direct reference to the discourse status of constituents. This includes information like whether the target constituent is lexically inferable from its antecedent, but potentially broader considerations as well, such as whether the meaning of the target constituent is made available by the nonlinguistic context. By contrast, the accommodation approach simplifies the grammar of deaccenting by consider-

ing only lexical identity with the antecedent, and moves broader discourse considerations to an independent pragmatic process.

Thus, the goal of this half of the dissertation is twofold. First, it constitutes a systematic survey of the prosodic behavior of accessible constituents in both production and perception. It addresses many of the empirical gaps highlighted above, assessing both whether naive speakers actually deaccent inferable material in production, and whether naive listeners find deaccented inferable material acceptable in perception. Second, it aims to adjudicate between the grammatical and accommodation approaches to the licensing of deaccenting for inferable material by assessing whether the production and perception data are more consistent with one approach than the other.

To briefly preview the experimental approach, the next chapter opens with a production study. In this experiment, participants read sentences of the form *subject-verb-object and subject-verb-object* (*SVO and SVO*). The first SVO clause was manipulated so the second-clause verb could be new, inferable, or repeated. The recorded productions were analyzed phonetically to determine whether inferable verbs were accented or deaccented, or whether the results exhibited an intermediate phonetic status or variability of prosodic realization. This analysis was also validated with a perception study eliciting naive listeners' assessments of whether the recorded words were emphasized or not emphasized.

Following the production study, the subsequent chapter presents a series of perception studies. These experiments feature manipulated recordings generated using the design of the production study. *SVO and SVO* sentences had their two clauses cross-spliced so that verbs that were canonically accented or deaccented in production could appear in an environment making them new, inferable, or repeated. This represents a novel systematic investigation of whether listeners find deaccented inferable material acceptable, and informs the theory of deaccenting licensing separate and apart from the contribution of the production study.

Together, the experimental results indicate that deaccenting of antecedent-nonidentical material is more restricted than has been proposed in the literature. In production, speakers

reliably accented inferable/accessible verbs rather than deaccenting them. In perception, sentences with deaccented inferable verbs were rated as relatively infelicitous when rated out of the blue. However, the addition of a discourse context supporting an identical reading for the antecedent and inferable target verbs salvaged ratings somewhat, while adding the presupposition trigger *too* brought naturalness ratings of deaccented inferable verbs up to the level of deaccented repeated verbs.

These results represent a novel empirical characterization of deaccented inferable material, namely that it requires substantially more support from the broader context (that is, beyond the target and the antecedent) than has been proposed in the literature. The results also hold theoretical significance in that they show clear empirical differences between antecedent-identical (repeated) and -nonidentical (inferable) material in both production and perception. This is predicted by the accommodation account of deaccenting under nonidentity, but not the grammatical account. Thus, the dissertation concludes that the accommodation account is on the right track, and suggests that other theories must be adjusted to capture this empirical observation.

In preparation for these experimental investigations, the section immediately below reviews the prior literature on the phonetics of accentuation and deaccentuation. This serves primarily to motivate the phonetic analysis performed on the production data, which measures intensity, f_0 , and duration and takes these as sufficient to diagnose whether a constituent is accented or deaccented.

It should be noted that there is also a large literature on the phonology of accent that is not treated in detail here. The task of characterizing accent (or lack thereof) in terms of discrete mental categories is, of course, highly complex (Pierrehumbert, 1980; Pierrehumbert and Hirschberg, 1990; Ladd, 1996). This line of research has culminated in a standardized system for annotating phrasal accent in English called Tones and Break Indices, or ToBI (Silverman et al., 1992). This annotation scheme, added to the phonetic analysis, would be a useful tool for investigating the accent status of inferable material, and was employed by

Chodroff and Cole (2019) in their study of nouns. It is not employed or discussed in detail here merely to limit the scope of the dissertation.

The section following the phonetic background reviews prior theoretical treatments of deaccenting under nonidentity – the focus strategy, the givenness strategy, and the accommodation strategy – in more detail. This does not constitute an exhaustive treatment of any of the relevant concepts, particularly focus or givenness. Rather, the review concentrates specifically on models that have overtly mentioned deaccenting of inferable, accessible, or otherwise antecedent-nonidentical material. Resolving the interfaces of deaccenting with these larger concepts is a complex, but important, task that lies outside the scope of the dissertation.

5.1 Diagnosing emphasis

In the next chapter, this dissertation investigates the phonetic realization of discourse-accessible constituents in American English. The analysis presented will provide a preliminary characterization of the prosody of accessible material, a topic that has received very little systematic phonetic investigation. Further, by drawing phonological conclusions about the emphasis status of these constituents, the dissertation will investigate the extent to which the prior impressionistic characterization of the prosody of accessible material is empirically correct, and by extension, whether existing syntactic-semantic accounts of emphasis licensing need to be revised.

This section reviews the prior phonetic literature on prosodic prominence more generally. The next chapter opens with a production study investigating whether participants realize constituents of various levels of discourse accessibility as accented or deaccented. Following this are multiple perception studies that manipulate both accent status and discourse accessibility. As such, it is desirable to identify a set of quantitative acoustic variables that can provide an adequate description of the emphasis status a given constituent.

The upshot of this section is the selection of a set of three phonetic variables – mean

intensity, mean f_0 , and duration, as measured on syllabic nuclei – will allow for a sufficient characterization of whether the source constituent is accented or deaccented. At the same time, the review notes some limitations of this analytical approach, namely that it excludes some known acoustic correlates of prosodic prominence, glosses over subtleties in the realization of the correlates that are measured, and ignores interactions between adjacent segments and between adjacent prosodic units that are known to affect the specific realizations of the correlates of interest.

5.1.1 *Correlates of accent and correlates of stress*

Many English speakers, even those naive to the study of prosody, share strong intuitions about the role and realization of accent in the linguistic system. One intuition is that there is a principled connection between the role of a constituent in the discourse and whether it is realized with or without a pitch accent. In particular, material that is new to the discourse is generally realized with a pitch accent, and material that has been invoked recently is realized without a pitch accent. This intuitive mapping has largely been borne out in more systematic research (Halliday, 1967; Pierrehumbert and Hirschberg, 1990), although there are some complications that will be discussed further below.

Related to this intuition is an implicit understanding of what accent “is” in acoustic terms. English speakers might have the impression that emphasized material is associated with higher pitch, louder realization, or longer duration relative to material that is not prosodically prominent. There is a large body of research in phonetics that has focused on evaluating these intuitions, testing whether these and other acoustic correlates reliably pattern with phonological impressions of prominence and, more broadly, the information structural considerations that are thought to license prominence.

The phonetic literature has revealed reliable correlations between several acoustic variables and impressionistic assessments of prosodic prominence in English. These variables include:

- **f_0 :** Prominent material is associated with high or rising pitch excursions, whereas non-prominent material lacks such excursions. (Fry, 1955, 1958; Lieberman, 1960)
- **Intensity:** Prominent material is relatively louder, and non-prominent material is relatively quieter. (Fry, 1955, 1958; Lieberman, 1960; Beckman, 1986)
- **Duration:** Prominent constituents last longer than string-identical constituents that are not prominent. (Fry, 1955, 1958; Adams and Munro, 1978; Isenberg and Gay, 1978; Cutler and Darwin, 1981)
- **Voice quality:** Creaky voice is commonly associated with low pitch accented material and phrase boundaries, although its use varies widely by speaker and interacts with other prosodic characteristics such as word stress. (Catford, 1977; Wells, 1982; Henton and Bladon, 1988; Garellek, 2015; Chodroff and Cole, 2019)
- **Vowel quality:** Non-prominent vowels may receive a “reduced” realization, meaning they collapse toward a schwa-like realization, which can be measured using the F1-F2 formant space. Prominent vowels tend to receive a “fuller” realization, with F1-F2 patterns matching “canonical” realizations of the vowel phoneme. (Fry, 1965; Beckman and Edwards, 1994; Rosner and Pickering, 1994; Campbell and Beckman, 1997)

Numerous studies have sought to determine the relative importance of each of these cues to listeners’ perception of prosodic prominence (Fry, 1958; Flege and Bohn, 1989; Rosner and Pickering, 1994; Sluijter and van Heuven, 1996; Zhang and Francis, 2010). A number of preliminary investigations indicated that f_0 was the most important acoustic correlate for determining whether a constituent would be perceived as prominent (Fry, 1958; Bolinger, 1958; Morton and Jassem, 1965; Beckman, 1986).

However, subsequent work has shown that the relationships between the phonetic variables are more complex. For example, where some research has suggested that f_0 outranks intensity as a correlate of emphasis status (Rietveld and Gussenhoven, 1985; Terken, 1991),

other studies have shown the opposite (Kochanski et al., 2005). Furthermore, these two correlates are known to co-vary with one another in production (Hirano et al., 1969; Gramming et al., 1988). This tendency has been linked to the physiological processes involved in adjusting both pitch and intensity, but substantial variation between speakers has also suggested that physiology alone is not responsible for this connection (Tilsen, 2016).

More generally, the correlates of prominence have been shown to vary both directly and inversely with respect to one another (Gussenhoven, 2002; Rosenberg and Hirschberg, 2006; Fletcher, 2010; Rathcke, 2017). This means that it is important to examine several acoustic variables at one time, since examination of a single cue may obscure the actual phonological facts of prominence if that cue happens to be downweighted in a particular task or by a particular speaker (Breen et al., 2010). The picture is further complicated by the observation that the relative importance of each phonetic cue is likely parameterized differently by different languages (Morton and Jassem, 1965; Beckman, 1986; Sluijter and van Heuven, 1996; Dogil and Williams, 1999) and even by different dialects of the same language (Smith and Rathcke, 2020).

Much of the discussion above has appealed to a general notion of prosodic “prominence,” but it is important to note that this concept actually cuts across two related, but distinct, phenomena: lexical stress and phrasal accent. It is well known that English, like many languages, distinguishes between “stressed” syllables, which are relatively more prominent prosodically, and “unstressed” syllables, which receive a reduced realization. By contrast, accent is driven by higher-level considerations such as manipulations in information structure and the presence of boundaries between various prosodic phrases (Sluijter and van Heuven, 1996).

Several of the studies cited above, particularly earlier ones, were primarily concerned with identifying the correlates that were associated with listeners’ perception of which syllables were *stressed*, in the lexical sense, rather than which ones were *accented*. Since it is accent that is of interest in this dissertation, it is worth considering the extent to which the acoustic

correlates of accent are distinct from those of stress.

Sluijter and van Heuven (1996) were perhaps the first to note a critical conflation of stress and accent in much of the early work on prosodic prominence. Several of these studies elicited words in isolation or at the end of a carrier phrase. However, Sluijter and van Heuven noted that words in these positions are associated with independent prosodic boundaries that cause them to be accented, meaning that the prior conclusions about the correlates of syllabic stress may have inadvertently included information about accent, as well.

Sluijter and van Heuven conducted a study aimed at disentangling the acoustic correlates of stress and accent. They systematically varied whether the words they elicited were produced with or without focus, allowing them to determine which cues are associated with accent (under the assumption that focused words carried phrasal accent) and which were associated with stress. They found that f_0 and intensity were actually very weakly predictive of syllable stress, while other cues like vowel quality and voice quality remained reliable predictors of judgments of stress. By contrast, f_0 and intensity were highly predictive as correlates of phrasal accent, and the presence of focus also correlated with duration, as words were roughly linearly dilated when realized with focus compared to when they were realized without. Subsequent studies have confirmed the association of similar cues as correlates of accent in particular; for instance, Chodroff and Cole (2019) found significant differences in voice quality, intensity, and duration based on the constituent's discourse status (given vs. new or contrastive).

This background helps outline the type of phonetic analysis that can be used to classify constituents as accented or deaccented in the following chapter. It goes without saying that all of the acoustic variables analyzed should be known correlates of accent. Furthermore, given the somewhat nondeterministic mapping between accent and individual acoustic correlates, it would be prudent to examine several correlates rather than focusing on a single variable.

The analysis presented in the next chapter focuses on three acoustic variables measured

on vocalic nuclei: mean intensity, mean f_0 , and duration. Previous studies have shown each of these to be reliable correlates of accent specifically, as analyzed separately from lexical stress. They can also be measured more or less automatically with a script given a properly annotated audio file.

The major correlates of prominence that are left out of this analysis are voice quality and vowel quality. Voice quality was excluded primarily in the interest of time, as it is less clear that this can be measured automatically using a script, although there are known phonetic correlates of modal versus creaky voice. Voice quality will be considered during data processing for the f_0 analysis, so some impressionistic data on the distribution of creaky voice in the data will be available.

Vowel quality was excluded primarily because the identity of the critical vowel is not held constant across items, so it is not immediately clear how to make principled comparisons between conditions. (One possibility might be to compare difference scores of F1-F2 measures between conditions for the same participant, since in the production study, each participant produced every item in every condition).

Voice quality and vowel quality are less reliable correlates of emphasis than intensity, f_0 , and duration (Sluijter and van Heuven, 1996), so it seems reasonable to expect that the experimental paradigm will appropriately characterize the accent status of the target constituents even given these exclusions. As will be seen in the next chapter, the results using the three selected correlates are both internally consistent and with theoretically motivated expectations for the acoustic realization of the experiment's baseline discourse-new and discourse-given conditions. This suggests that intensity, f_0 , and duration are sufficient for an initial characterization of the accent patterns of discourse-accessible material, although further analysis of other correlates of prominence could certainly yield novel results in the future.

In light of the prior literature on accent in English, we opted to focus on three correlates of accent in our phonetic analysis below. These are the mean f_0 of the target word's stressed

vowel, the mean intensity on the vowel, and the vowel's duration. This is, of course, an oversimplification of prosodic production. This analysis ignores several potentially relevant variables, such as the placement of f_0 peaks within a syllable, voice quality, and vowel quality, and does not take into account prosodic information on other segments in the target word. The stimuli are also not controlled for certain factors that are known to affect prosodic realization, such as the voicing of consonants adjacent to the target vowels.

Despite these shortcomings, we expected the measures we chose to provide an adequate preliminary picture of the production facts for given and accessible material based on previous work on the relative importance of different phonetic cues (e.g., Sluijter & van Heuven 1996). The stimuli are controlled across items to the extent possible, with the number of syllables before and in the target clause held constant for all items and a constant stress pattern in the critical sentences themselves. In addition, the design was balanced across conditions, with each participant producing string-identical critical clauses in all possible conditions, which we expect will help reduce some of the variability associated with the segmental environment of the target vowels. The internal consistency of the results for discourse-new versus -given constituents, whose canonical prosodic realizations are better understood, suggest that our paradigm provides a fairly robust characterization of the prosody of the elicited sentences. However, we acknowledge that future work examining these correlates in more detail, as well as work looking at correlates that we did not study, could potentially yield different results.

5.1.2 *Complications of prosodic realization*

The acoustic analysis outlined above oversimplifies the phonetic nature of prosodic prominence in multiple ways. It is worth considering these limitations as future research taking them into account may yield qualitatively different results from those reported in the next chapter.

First, as discussed above, the analysis leaves out some known correlates of prosodic prominence such as voice quality and vowel quality. Relatedly, the analysis of the production study

does not feature phonological coding of pitch accent types using a theoretical framework such as ToBI (Silverman et al., 1992). Rather, the analysis “crowdsources” phonological judgments by collecting a large number of native-listener intuitions about whether particular words in the recorded productions are emphasized. Importantly, analysis of phonological judgments of prominence yielded a key difference between the present study and Chodroff & Cole’s 2019 study of the production of discourse-accessible nouns. As will be seen in the next chapter, the present study finds no reliable phonological differences between discourse-new and discourse-accessible verbs. By contrast, Chodroff & Cole’s study involved trained pitch accent annotation, and they found a significant difference in the proportion of times accessible versus new or contrastive verbs received a high or rising pitch accent.

Second, the correlates that will be used somewhat simplify the actual critical facts of the phonetic realization of accent. For instance, considerable work has been done detailing the acoustic nature of pitch excursions that are relevant to the classification of pitch accent categories, including the timing and shape of the excursions within a syllable (e.g., Pierrehumbert and Hirschberg, 1990). Because analyzing every syllable in the production study would be highly labor intensive, this study uses mean f_0 across the full nucleus as a proxy measure to determine which syllables are relatively more or less likely to have a meaningful high or rising pitch excursion.

It is possible that the measures of mean intensity and vocalic duration similarly obscure the true prosodic nature of the corresponding constituents. It may be the case that intensity exhibits similar contours over the course of a vowel to f_0 , whereas in the present study, only mean intensity for the entire vowel will be measured. Likewise, Sluijter and van Heuven (1996) note that accented words undergo a roughly linear time dilation compared to unaccented words. The present methodology of measuring only vocalic duration may obscure some meaningful prosodic information on other segments in the target words, e.g., the duration of approximant, nasal, or fricative consonants.

For the bulk of the analysis, the next two chapters will largely assume a one-to-one

mapping between more familiar discourse statuses, such as “given” and “new,” and accent categories like “unaccented” and “accented,” respectively. However, it is worth noting that this clean mapping has already been problematized in the literature. While the given-unaccented/new-accented mapping is a robust trend, it is not universally adhered to. Several studies have shown the opposite mapping, for example, with given information being realized with a high or rising pitch accent (Bard and Aylett, 1999; Ito et al., 2004; Chodroff and Cole, 2019). This has led some prosody researchers to propose that the mapping should be treated as probabilistic, rather than deterministic (Chodroff and Cole, 2019). The present analysis will largely ignore this; although there are likely to be substantial individual-level and trial-level differences in both the production and perception of accent, these specifics will be treated as outside the scope of the present study.

There are several additional factors that affect the realizations of the phonetic correlates that have been selected for analysis. One major confound for prosody research is the observation that constituents in phrase-final position are generally lengthened, possibly separate from the effect of any non-boundary accent the constituent carries (Klatt, 1975; Umeda, 1975; Wightman et al., 1992). This effect is partially controlled for in the present production study, as the most direct analytical comparisons will be made between constituents in the same position. However, it is possible that there is an interaction between accent status and boundary-associated lengthening that will somewhat affect the results in the duration analysis.

There is a variety of studies indicating that many acoustic correlates of prominence, including intensity, duration, and vowel quality, can be affected by the surrounding segmental environment (House and Fairbanks, 1953; Peterson and Lehiste, 1960; House, 1961; Chen, 1970; Raphael, 1972). Likewise, it is well known that segmental features such as adjacent consonant voicing can affect the f_0 realization of a vowel (House and Fairbanks, 1953, i.a.). Since the present design compares identical lexical items across conditions, and participants produce each item in each condition, this is partially controlled for. However, it is possible

that the segmental contexts of the target vowels will affect the prosodic measurements of interest in unexpected ways.

Finally, it has been established that acoustic correlates of prominence on a particular syllable can affect syllables other than the target, as well. For instance, the timing of a pitch accent within a syllable is affected when an adjacent syllable contains a pitch accent compared to when the adjacent syllable is unaccented (Silverman and Pierrehumbert, 1990). Furthermore, acoustic markers of emphasis on one syllable have been shown to spread to adjacent syllables (Turk and White, 1999; Cambier-Langeveld and Turk, 1999; Turk and Shattuck-Hufnagel, 2000; Cho and McQueen, 2005; Cho and Keating, 2009). These considerations are potentially relevant for the current study, as the design directly manipulates whether adjacent constituents will or will not both be accented, and the analysis assumes that correlates of emphasis observed on a particular syllabic nucleus indicate accent on that syllable, as opposed to effects of the prosodic marking of adjacent syllables.

In light of these potential confounds, it will no doubt be worthwhile to revisit the present production data in the future to examine the prosodic realization of individual tokens, taking into account finer-grained techniques for diagnosing prosodic prominence as well as the broader prosodic environment of each word. Nevertheless, given the high degree of control built into the experimental design, as will further be explained in the next chapter, it seems fair to assume that the analytical protocol selected here will provide an adequate preliminary picture of the phonetics of accessible material. Further, the results of the production study are highly internally consistent and map well onto naive listeners' judgments of emphasis, suggesting that these narrower considerations did not have a substantial effect on the qualitative pattern of results.

5.2 Theories of deaccenting licensing

This section reviews several theories of the assignment of prosodic emphasis, concentrating on work that explicitly generates or makes predictions about cases of deaccenting under

nonidentity. As discussed above, this work is discussed in two groups. First, there is work proposing that deaccenting is grammatically licensed when it is given or inferable in the discourse. This includes work giving structural, focus-theoretic characterizations of deaccentable material calculated with reference to a linguistic antecedent, as well as work that assumes holistic, strategic marking of discourse-givenness. Second, there is work that explicitly appeals to formal identity in the grammar of deaccenting, with nonidentical cases derived through an extragrammatical operation. The former approach is noted to skew in favor of “grammatical” licensing of deaccenting under nonidentity, meaning that such deaccenting is automatically ruled in by the grammar. However, it should be noted that the theories can be highly complex, and some leave room for additional extragrammatical processes beyond the straightforward licensing mechanism. By contrast, the latter group of accounts takes an “accommodation” approach to deaccenting under nonidentity, marking these cases as ungrammatical but nevertheless acceptable.

5.2.1 *The grammatical approach*

It is well established that givenness and accessibility have complex effects on the ways speakers choose to encode meaning, although much of the prior work on this domain has concentrated on referring expressions (e.g., Ariel, 1988, 1990; Gundel et al., 1993). Several different dimensions of givenness have been argued to impact linguistic encoding (Prince, 1981), including whether the speaker assumes that the hearer is broadly familiar with some concept (Kuno, 1972; Stalnaker, 1974; Clark and Haviland, 1977; Clark and Marshall, 1981), whether the speaker assumes the concept is active in the listener’s consciousness or otherwise predictable (Chafe, 1976, 1994), and whether the speaker wishes to signal that the concept is important or unimportant to the discourse (Halliday, 1967; Kuno, 1972). Later approaches have sought to unify some or all of these notions, such as Prince’s (1992) analysis of information along two dimensions, hearer-new/hearer-old and discourse-new/discourse-old.

Chafe (1974) argues that whether a concept is currently “active” in the hearer’s consciousness is the most important (or perhaps the sole) predictor of whether it can be deaccented. He notes that material need not have been used in the previous discourse in order to be deaccentable, as in (65).

(65) [A man walks by in a way noticeable to both speaker and hearer]

I KNOW *that guy*.

(Chafe, 1974)

At the same time, mere knowledge of a concept or referent is not generally sufficient to license deaccenting, as shown in (66), where the listener presumably already knows there is a place called Japan. This suggests that concepts generally need to be evoked in the linguistic or nonlinguistic context before becoming deaccentable.

(66) I’d like to go to JAPAN.

(adapted from Chafe 1974)

As it turns out, accent assignment is somewhat more complicated than a straightforward mapping between conscious activation and deaccentuation would suggest. The accounts reviewed in this subsection aim to characterize the notion of givenness that is relevant to the licensing of deaccentuation, with the assumption that it is givenness, rather than lexical identity, that licenses deaccenting in the grammar. The approaches taken lie on a spectrum from ones that take a highly structured approach to defining givenness in terms of formal relationships with an antecedent to ones that treat givenness marking as a largely strategic move made by speakers to help hearers make sense of the discourse. The individual accounts are difficult to order along this continuum, so they are discussed in roughly chronological order below. It should be noted here that this is not an exhaustive review of either focus theory or work on givenness, but rather a survey of work in these domains that makes explicit theoretical arguments about the licensing of deaccenting for constituents without identical correlates in the antecedent.

Ladd: Focus, co-reference, and “shifters”

Many early accounts of prosodic prominence sought to explain accent placement entirely in syntactic terms. Chomsky and Halle (1968) proposed a nuclear stress rule that assigned English phrasal accent on the basis of constituent structure. After the assignment of word stress, this rule cyclically strengthened accent on the rightmost element of each major constituent, which generally reproduced intuitions about “default” stress in English declarative sentences. Much subsequent work focused on refining the nature of this rule while maintaining that accent assignment was fundamentally driven by syntax (Bresnan, 1971; Berman and Szamosi, 1972; Lakoff, 1972; Stockwell, 1972).

Bolinger (1972) objected to this characterization on the basis of examples like (67).

- (67) a. My mother is EATING.
b. #My mother is COMING.
c. My MOTHER is *coming*.

(67a) represents a fairly typical declarative sentence with a nuclear pitch accent on the rightmost constituent. Despite having (superficially) identical syntax, however, the corresponding prosodic realization of (67b) is generally not felicitous out of the blue. Rather, a realization like (67c) is more typical. Bolinger cited such examples as evidence that syntax alone is not the sole or even primary factor determining the placement of accent. Instead, he argued that accent is driven by strategic discourse goals on the part of the speaker, such as the intent to signal the “emotional weight” of a particular constituent. However, the exact dimensions that were proposed to correspond with presence or absence of accent were not clearly defined (Schmerling, 1976).

Ladd (1980) sought to combine the intuitions of these two approaches – that there is a principled mapping between syntactic structure and accent placement on the one hand, but that this mapping is affected by the structure of the discourse and the speaker’s goals on the other. His approach builds on prior work on *focus*. Intuitively, focus marks the part of a

sentence that contributes new, contrastive, or otherwise remarkable information relative to the rest of the material.

Often, focus is diagnosed by examining the types of questions a sentence can felicitously answer. In many positions, a nuclear pitch accent gives rise to a reading of “narrow” focus, meaning that only a small constituent containing the accented expression is focused. For example, in (68), the response with a pitch accent on MARY is a felicitous reply to the *Who?* question because MARY is in narrow focus. However, it is not a felicitous response to the question in (69).

(68) A: Who bought a book about bats?

B: MARY *bought a book about bats*.

(69) A: What’s been happening?

B: #MARY *bought a book about bats*.

By contrast, much early work on focus noted that a pitch accent in a particular position, namely the rightmost or most embedded constituent in the sentence, can give rise to a much wider range of readings (Halliday, 1967; Chomsky, 1971; Jackendoff, 1972). With pitch accent in this position, focus can be interpreted as narrow or broad, meaning it applies to a larger constituent or the sentence as a whole, or unspecified. For instance, in (70), the response with a nuclear pitch accent on BATS is a felicitous reply to any of the questions, indicating an ambiguity as to whether [bats], [about bats], [a book about bats], [bought a book about bats], or the full sentence is in focus.

(70) a. A: What did Mary buy a book about?

b. A: What kind of book did Mary buy?

c. A: What did Mary buy?

d. A: What did Mary do?

e. A: What’s been happening?

B: Mary bought a book about BATS.

(Selkirk, 1995)

Ladd's proposal was to treat accent assignment entirely in terms of focus rather than to treat focus as a phenomenon that merely interacted with a separate set of rules assigning accent according to syntax. Essentially, Ladd reformulated the rule for assigning accent so that it made reference to focus domains instead of syntactic constituents in general. So, instead of assigning a nuclear pitch accent to the rightmost accentable item in a sentence, it would be assigned to the rightmost accentable item within a focused constituent.

Importantly, however, Ladd's proposal allowed for some flexibility in assigning accent beyond his incorporation of the focus domain. Expressions are allowed to be *deaccented* separate and apart from strict focus considerations if they stand in an appropriate connection to material from earlier in the discourse, or in the context at large. For instance, in (71), *the measure* is deaccented because it co-refers with *the bill*.

- (71) A bill was sent to Congress today by President Carter which would require peanut butter sandwiches to be served at all government functions. At a press conference today, a group of senators led by Republican Barry Goldwater DENOUNCED *the measure*.

Co-reference is not mandatory for deaccenting, however; (72) shows that deaccenting is possible even when a concept is available in the discourse, even if the expression and its antecedent do not co-refer.

- (72) A: Has John read Slaughterhouse-Five?

B: No, John doesn't READ *books*.

Deaccenting can even be supported by broader contextual considerations. For instance, in (73), the response in (73a) might be appropriate among general audiences, for instance a parent discussing their son's dissertation with an interested friend. By contrast, Ladd sug-

gests that (73b) is appropriate among linguists, where it is presupposed that the dissertation would be about language in some way.

- (73) A: What's his dissertation about?
- a. B: Something about LANGUAGE *acquisition*.
 - b. B: Something about *language* ACQUISITION.

Ladd essentially exempts deaccented constituents from the general accent assignment rule. If a nuclear accent would normally land on a constituent that is chosen for deaccenting, it must be located elsewhere (usually by moving to the left). Thus, Ladd's approach allows for clear and predictable effects of syntactic structure on accent placement, but also respects the observation that the context of utterance has an effect on prosody even in commonplace sentences (Bolinger, 1972; Schmerling, 1976).

Turning to the question of what constituents can be deaccented, Ladd does not spell out the requirements in systematic detail. Rather, he proposes understanding deaccenting in terms of Jakobson's (1971) notion of a *shifter*. He suggests that deaccenting a constituent is a signal that the constituent now relies on the broader context for its full interpretation. For instance, deaccenting *the measure* in (71) signals that there is some prior entity in the discourse that is relevant to the interpretation of this constituent. A speaker might choose to deaccent rather than using another type of shifter, namely a pronoun like *it*, since deaccenting allows for the maintenance of lexical content in the constituent that can help guide interpretation. In essence, deaccenting becomes a middle ground that simultaneously contains non-vacuous meaning of its own but also signals a relationship to the broader context.

Thus, Ladd schematizes deaccenting primarily as a strategic choice made by speakers to help guide listeners to the correct interpretation of their utterances. He does not provide strict criteria for what constituents can be deaccented. This leaves the door open for several types of deaccenting under nonidentity. The examples in (71) through (73) demonstrate at

least three types of deaccentable constituents: those that co-refer with an antecedent, those that can be inferred from an antecedent (in this case, by a superset relation), and those that are made available by broader knowledge of the context of utterance or about the world. Subsequent approaches to deaccenting would build on this by attempting to formalize the set of deaccentable constituents more systematically.

Selkirk, Rochemont, et al.: Givenness and c-construability

Selkirk (1995) summarizes an approach to emphasis licensing developed by Gussenhoven (1984), Selkirk (1984), and Rochemont (1986). Recall that a core phenomenon of interest in the study of focus is the fact that certain accent configurations are ambiguous with respect to the scope of focus. A sentence with a pitch accent in an appropriate position can be the answer to several different questions, as demonstrated again in (74).

- (74) a. A: What did Mary buy a book about?
 b. A: What kind of book did Mary buy?
 c. A: What did Mary buy?
 d. A: What did Mary do?
 e. A: What's been happening?

B: Mary bought a book about BATS.

(Selkirk, 1995)

The theory accounts for this data by positing a mechanism of *focus projection*. This framework requires that any prosodically prominent element (BATS in (70)) hosts a marker called *F*. Any F-marked node that is not dominated by another F-marked node is given the label FOC; the material dominated by FOC is interpreted as the focus of the sentence in the informal sense (i.e., the answer to a *wh*-question). The architecture described so far accounts for the reply in (74) as a response to (74a). BATS is necessarily F-marked, since it is prosodically prominent. On the reading where the reply responds to (74a), BATS is also interpreted as a FOC, with no other F-marked nodes dominating it.

To account for the other focal scopes that are possible for this sentence, two principles of *focus projection* are posited that allow for the percolation of F-markers and therefore for expansion of the material included in FOC. The two principles are (1) that F-marking of a phrasal head licenses F-marking of the phrase itself, and (2) that F-marking of a head's internal argument licenses F-marking of the head. Contra Chomsky and Halle (1968), Chomsky (1971), and Jackendoff (1972), the principles of focus projection make reference to argument structure instead of linear order. This is due to observations from languages with similar prosodic properties, e.g., German and Dutch, that the assignment of default prosodic prominence is driven by argument structure rather than linear order (Selkirk, 1984, 1995). Note that projection of the F-marker is merely licensed, not required; F-markers are postulated to project only as required to make an utterance felicitous given the required scope of focus.

Thus, the reply in (74) is marked as a felicitous response to the remaining questions by means of focus projection. Since BATS is the internal argument of *about*, F-marking of *about* is licensed, and since *about* heads the PP *about* BATS, the full PP can be F-marked. If F-marking stops here, then PP is a FOC, and the utterance felicitously responds to (74b). This process can be iterated all the way up to full sentence focus, accounting for the congruence between the reply and all of the questions in (74).

This approach correctly predicts the ambiguity – or lack of ambiguity – of the focus interpretation of a variety of constructions. Consider again (75) and (76).

(75) A: Who bought a book about bats?

B: MARY *bought a book about bats*.

(76) A: What's been happening?

B: #MARY *bought a book about bats*.

In both replies, MARY must be F-marked because it is prosodically prominent. In (75), MARY is a FOC, which correctly results in subject focus, a felicitous reply to the question. By contrast, the question in (76) requires sentence focus in the reply. However, the principles of focus projection do not allow for a larger FOC than MARY in this case, since Mary is

neither the internal argument of any other element nor the head of a larger phrase. Thus, this prosodic configuration is infelicitous as a response to the question in (76).

Selkirk (1995) notes three properties linking F-marking to discourse status in the focus projection system. The first is that non-F-marked material must necessarily be interpreted as given in the discourse. For instance, in (75) above, *bought a book about bats* is given in the antecedent, and none of these elements are F-marked in the reply. The second is that F-marked material that is not a FOC must be interpreted as new. This is demonstrated by the reply in (77), where there is an additional prosodic emphasis on the subject.

(77) A: What's been happening?

B: MARY bought a book about BATS.

(Selkirk, 1995)

The question in (77) requires sentence focus in the reply, and the full sentence is a FOC due to projection from the obligatory F-mark on BATS. However, since MARY is also prominent, it must be F-marked, but does not contribute to the placement of FOC via projection. Selkirk argues that such a prominence is only possible when the constituent is interpreted as new.

Finally, Selkirk notes that a FOC can be interpreted as given or new in the discourse. The intuition here is that a constituent can be given in the discourse, but separate focus requirements in the discourse can force it to host an F-mark in order to generate a FOC in the correct position, as shown by (78).

(78) A: After all that talk about bicycles yesterday, do you know what Kim bought today?

B: *Kim bought a* BICYCLE.

(Büring, 2016)

In (78), BICYCLE is given in the antecedent. However, the question requires a reply with object focus, so the object must be a FOC in the reply. To achieve this, BICYCLE must host an F-mark, making it a FOC. This demonstrates that being a FOC in Selkirk's system is not incompatible with being discourse-given.

So, in Selkirk’s framework, deaccented elements are licensed when two conditions are met: the element must be given, and it must not separately be in focus. Relevant to this dissertation, then, is the question of what counts as given. It is straightforward that elements that have an identical correlate in an antecedent utterance are given, and the examples Selkirk (1995) uses to demonstrate the role of givenness generally take this form. Of interest here are elements that do not have an identical antecedent correlate.

Within the focus projection framework, this question is taken up primarily by Rochemont (1986), following Culicover and Rochemont (1983). Rochemont proposes that a constituent must be a focus if and only if it is not *c-construable*. Foci receive pitch accents, whereas non-foci, which by extension must be c-construable, are not accented. What actually counts as c-construable is a complex issue, and there are a number of ways an element can be c-construable.

First, Rochemont proposes that a constituent is *directly* c-construable if it has a semantic antecedent in the discourse – that is, there is an antecedent string that “formally or informally” entails the string of interest. Formal entailment refers to logical entailment; Büring (2016), summarizing the types of antecedents that are sufficient for the focus projection approach, lists synonyms, hyponyms, and coreferential expressions as possible antecedents for a non-F-marked constituent. Informal entailment, on the other hand, refers to information a hearer might reasonably conclude from the antecedent despite it not being logically entailed. Thus, deaccenting of non-identical constituents is licensed via direct c-construability when there is a suitable linguistic antecedent, as in (79).

- (79) Harry wants a Volkswagen, but his wife would prefer an AMERICAN *car*.
(Ladd, 1980)

In this case, *car* does not have an identical correlate in the antecedent. However, *car*, or *German car*, is entailed by *Volkswagen* in the antecedent, making the mention of *car* in the second clause c-construable and licensing its lack of accent.

Rochemont also proposes that a constituent can be directly c-construable if it has a semantic antecedent in the broader discourse, even in the absence of a linguistic antecedent, as in (80).

(80) [Hearer cocks their head to one side as if listening for a faint or distant noise.]

Speaker: I *heard it*, TOO.

(Rochemont, 1986)

In this case, the hearer tilting their head makes it salient that they think they heard a noise, and this information can serve as a semantic antecedent for *heard* in the speaker's utterance. Since *heard* has a semantic antecedent in the discourse, it is directly c-construable, and it can be realized without a pitch accent.

In addition to direct c-construability, it is possible for a constituent to be *indirectly* c-construable. Rochemont lists two cases of deaccenting he considers to be due to indirect c-construability. The first involves of “verbs of appearance,” which often appear with no pitch accent even though they are not given in the discourse, as in (81).

(81) A LETTER *arrived for you yesterday*.

(Rochemont, 1986)

Rochemont argues that verbs like *arrive* are lexically specified as c-construable, and as a result can be a deaccented non-focus even when they do not have an antecedent. He suggests that this “unremarkability” is part of the conventional meaning of such words, comparing lexically-specified unremarkability to Grice's (1975) conventional implicatures, such as the lexically specified contrastiveness that distinguishes but from *and*.

The second way Rochemont suggests a constituent can be indirectly c-construable is if a speaker and hearer take deaccented information to be unremarkable background information about one of the discourse participants, as demonstrated in (82).

(82) a. I ran into JOHN at the HOSPITAL this morning.

b. I ran into JOHN at the *hospital* this morning.

(Rochemont, 1986)

The realization in (82a) might be considered the default way of pronouncing this string. It is felicitous when the speaker's being at the hospital is not considered unremarkable background information. For instance, a speaker who works as a teacher and goes to the hospital only very infrequently might utter (82a).

By contrast, (82b) is felicitous if the discourse participants know that the speaker is often at the hospital and take this as unremarkable background information. For instance, the speaker might be a nurse who works in a hospital several times a week. In this case, the speaker's having been in the hospital this morning is unremarkable, which is sufficient to render *hospital* c-construable and license the lack of pitch accent. Rochemont compares this indirect c-construability to Grice's conversational implicatures, since they depend on the individual knowledge states of the discourse participants and are not inextricably linked to specific lexical items.

Rooth: Alternative semantics

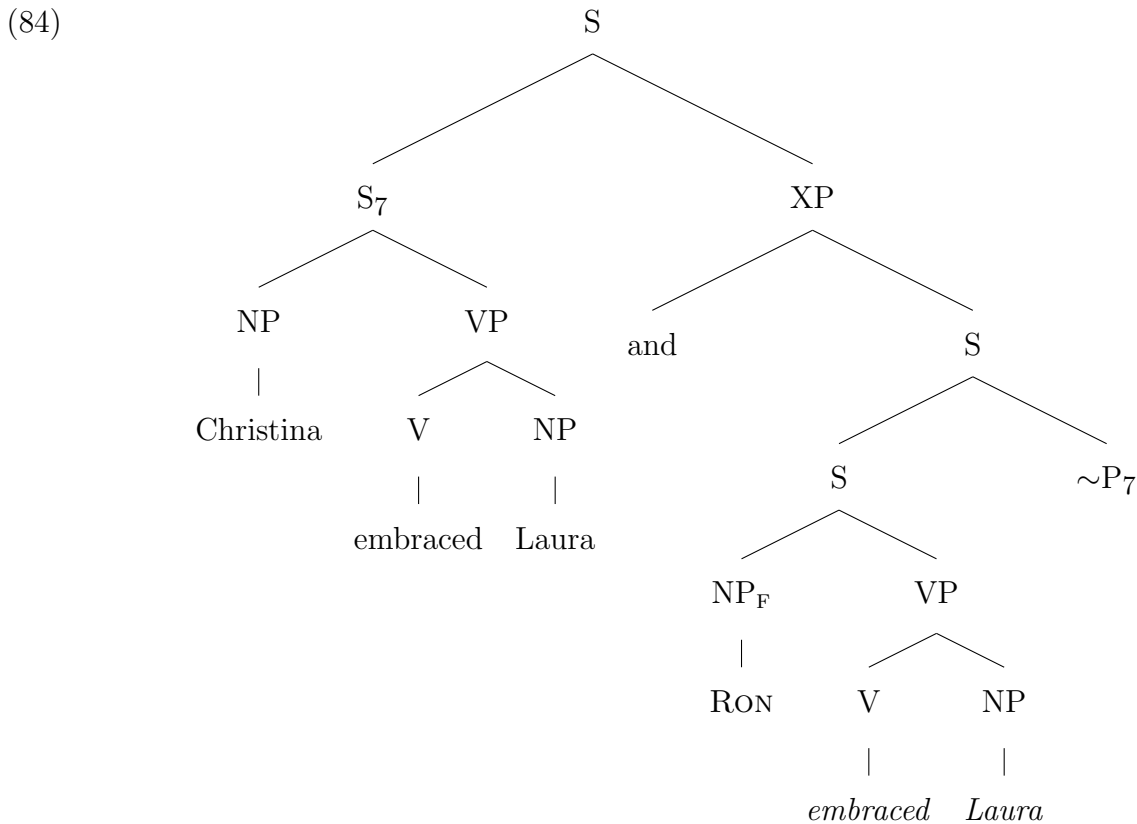
Rooth (1992b) presents a theory of the representation and interpretation of focus that builds on his earlier account of the compositional derivation of focus alternatives (Rooth, 1985). This theory uses a syntactic operator, which Rooth calls $\sim$, to retrieve the focus alternatives of syntactic constituents and make them compositionally accessible. $\sim$ takes two arguments, one overt and one covert. The overt argument is a syntactic constituent, the domain over which focus is interpreted. The covert argument is a free variable; informally, for a particular focus construction to be felicitous, there must be a contextually available value for the covert free variable that is a focus alternative of the overt argument of $\sim$.

In Rooth's system, prosodic emphasis is driven by the distribution of F-markers in a constituent. F-marked terminal nodes are prosodically emphasized, while non-F-marked nodes are deemphasized (deaccented). For a constituent to be either F-marked or non-F-marked, it must have an antecedent that is a focus alternative. For non-F-marked constituents, the

only focus alternative is the constituent itself. In this way, the system construes deaccenting as an operation of identity; within a focus domain, a constituent can only be non-F-marked, and therefore deaccented, if an identical meaning is made contextually available (but see below for cases of deaccenting under nonidentity).

As an illustrative example, consider (83), where in the second clause that *Ron* is contrastively F-marked and *embraced* and *Laura* are not. In Rooth's theory, this utterance might have the structure shown in (84).

(83) Christina embraced Laura, and RON *embraced Laura*.



In (84), the overt argument of $\sim$ is the second SVO clause, *Ron embraced Laura*. The covert argument is the free variable shown as P. According to Rooth's definition of $\sim$, this structure is subject to a presupposition that there is a contextually available meaning that is a focus alternative of the overt argument; informally, this means a structure of the form *x embraced Laura*. Since the first clause, *Christina embraced Laura*, is of the form *x embraced*

Laura, it is a focus alternative of the second clause and can be the contextually supplied value for the variable P, indicated by the coindexing of the first clause and P. Since the presupposition of $\sim$ is satisfied, this focus configuration is felicitous.

Rooth's system uses the framework of *alternative semantics* (Rooth, 1985; Kratzer, 1991) to model focus alternatives. In this framework, constituents have two semantic values – the *ordinary semantic value*, and the *focus semantic value*. The ordinary semantic value of X, indicated as $\llbracket X \rrbracket_{\mathcal{O}}$, is simply the denotation of X. The focus semantic value of X, indicated as $\llbracket X \rrbracket_{\mathcal{F}}$, which intuitively is the set of alternatives from which X was drawn.

Here, it will suffice to discuss the calculation of the focus semantic values in informal terms, but one formalization is given by Rooth (1985). When a constituent does not contain any F-marking, its focus semantic value is the singleton set containing the constituent's ordinary semantic value. When there is F-marking in the constituent, the focus semantic value can be generated by replacing any focused node with a variable of the same semantic type. The table in (85) shows some example constituents and schematizes their focus semantic values.

	X	$\llbracket X \rrbracket_{\mathcal{F}}$
	embraced <i>Laura</i>	embraced <i>Laura</i>
(85)	embraced _F <i>Laura</i>	R <i>Laura</i>
	embraced <i>Laura</i> _F	embraced y
	embraced _F <i>Laura</i> _F	R y

Thus, the focus semantic value of *embraced Laura* is the singleton set $\{\llbracket \text{embraced Laura} \rrbracket_{\mathcal{O}}\}$. The focus semantic value of *embraced_F Laura*, on the other hand, is the set of ordinary semantic values for all constituents consisting of a two-place predicate with *Laura* as the patient argument. Thus, $\llbracket \text{saw Laura} \rrbracket_{\mathcal{O}}$, $\llbracket \text{hugged Laura} \rrbracket_{\mathcal{O}}$, and $\llbracket \text{rebuffed Laura} \rrbracket_{\mathcal{O}}$ would all be in the focus semantic value of *embraced_F Laura*. Similarly, the focus semantic value of *embraced Laura_F* is the set of ordinary semantic values for *embraced* with any patient argument, and the focus semantic value for *embraced_F Laura_F* is the set of ordinary semantic values for any two-place predicate with any patient argument.

Rooth uses the notion of a focus semantic value to characterize the presupposition introduced by the $\sim$ operator. This presupposition can be satisfied under two conditions, depending on the semantic type of the overt argument of $\sim$. In the case that the overt argument denotes an individual, then the presupposition is satisfied if there is a contextually available value for the covert variable P such that P is an element of the focus semantic value of the overt argument. If the overt argument denotes a set, then the contextually available value for P must be a non-proper subset of the focus semantic value of the overt argument.

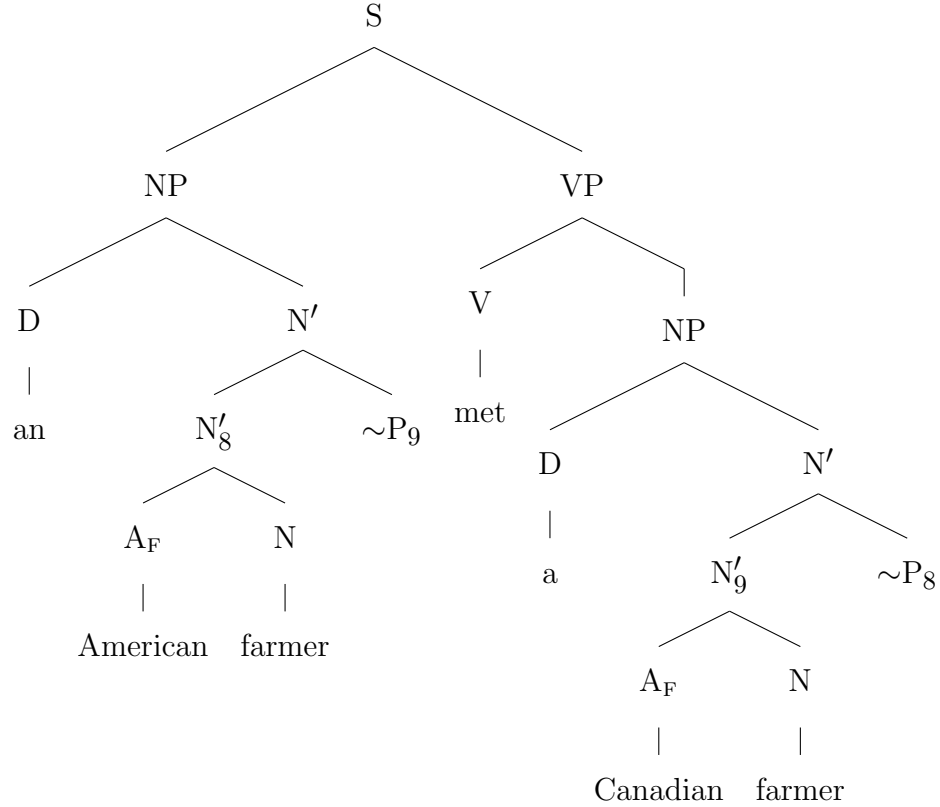
Thus, returning to (84), it is clear why the presupposition of $\sim$ is satisfied and the structure is licit. The presupposition is that there is some contextually available meaning that is a subset of the focus semantic value of *Ron_F embraced Laura*. Since *Ron* is F-marked, this is found by replacing *Ron* with a variable. Thus, the focus semantic value is any ordinary semantic value for a constituent of the form *x embraced Laura*. $\llbracket \textit{Christina embraced Laura} \rrbracket_{\mathcal{O}}$ has this form, so it is a subset of the focus semantic value of *Ron_F embraced Laura* and the presupposition is satisfied.

Rooth's use of the free variable P gives his account considerable flexibility in finding antecedents that license particular focus configurations. For instance, consider (86).

- (86) An AMERICAN *farmer* met a CANADIAN *farmer*. They had a fight about canola.
(Rooth, 1992b)

A theory of focus interpretation according to which focus configurations need to have overt linguistic antecedents would fail to rule this sentence in. The prosody of *American farmer* in the first-sentence subject seems to indicate contrastive focus, but this constituent does not actually contrast with any prior linguistic material. Rooth's theory allows the free variable argument of $\sim$ to freely look ahead in the discourse for its value, so this sentence has the structure shown in (87) (Rooth, 1992b).

(87)



Here, the free variable argument of $\sim$ in the subject takes the value of *Canadian farmer* in the object, while the free variable argument in the object takes the value of *American farmer* in the subject. This captures the intuition that the nouns *farmer* are mutually redundant, while the adjectives *American* and *Canadian* are contrastive, and that this configuration is prefigured by the lack of prosodic prominence on *farmer* in the subject.

Thus, Rooth's implementation of P as a free variable allows for relative flexibility in the satisfaction of the presuppositions on the use of $\sim$, including cases like the cataphora shown for P₉ above. This correctly rules in utterances like (86), but also suggests that P should be able to get its value in the same ways as other free variables, including, for instance, from the nonlinguistic context. This raises questions about whether the free variable approach may overgenerate in terms of the constituents in predicts can be deaccented, an issue that will be discussed more below.

Finally, while comparing the different identity requirements for deaccenting and verb phrase ellipsis, Rooth (1992a) notes a type of example, shown in (88), in which deaccenting

of a constituent is licensed by an antecedent that is not strictly a focus alternative of the constituent in question.

- (88) First John told Mary about the budget cuts, and then SUE_F *heard about them*.
(Rooth, 1992a)

In this case, only the subject of the second clause is focused, and so the focus semantic value of the clause is the set of all meanings of the form x *heard about the budget cuts*. Crucially, the only antecedent has the form y *told z about the budget cuts*, which is not strictly a focus alternative of the second clause.

Impressionistically, Rooth argues that the acceptability of (88) is due to the entailment relation implicit between the two clauses. *John told Mary about the budget cuts* is not a focus alternative for *Sue heard about the budget cuts*, but it does entail that Mary heard about the budget cuts, and *Mary heard about the budget cuts* is a focus alternative for the second clause. Rooth notes that this mechanism extends to examples where the relation is not driven by strict entailment, but rather by speakers' world knowledge ("presupposed axioms"), as in (89).

- (89) She called him a Republican, and then HE_F *insulted* HER_F .
(Lakoff, 1968; Rooth, 1992a)

In this case, *she called him a Republican* is not of the form x *insulted y* , and so it is not a focus alternative for the second conjunct. Moreover, it does not entail a proposition of the form x *insulted y* in any literal sense. However, Rooth suggests that as long as the speaker and hearer are willing to accept the axiom that calling someone a Republican constitutes an insult, then the focus configuration and indicated prosody in the second clause are felicitous via *implicational bridging*.

Rooth is agnostic about exactly how to account for the effect of entailment relations on deaccenting licensing. He suggests two possible implementations without discussing them in detail. The first is to account for the entailment and implicational bridging effect at

a level such as discourse representation, whereas the second is to directly incorporate an implicational relationship into the semantics of $\sim$.

van Deemter: Object-givenness and concept-givenness

van Deemter (1994, 1999) proposes an account of deaccenting where the mechanism of focus marking, and by extension accent assignment, responds directly to marking of givenness. In essence, any expression that is marked as given will be deaccented. However, van Deemter argues that there are actually several different types of givenness that can license deaccenting.

First, certain function words are specified as “minus lexically accentable” and can never be accented. These words are realized without accent by default. This assumption is shared, implicitly or explicitly, by most theories of accent assignment, and recalls previous proposals like Rochemont’s (1986) proposal that words like *arrived* are lexically specified as “c-construable.”

The second type of givenness is called *object-givenness*. Essentially, if an expression co-refers with an antecedent, then it is object-given and deaccentable. This can lead to deaccenting of antecedent-nonidentical constituents, as an expression only needs to co-refer with an antecedent to be deaccented. For example, in (90), *the town* and *Leiden* co-refer, so *the town* is object-given and can be deaccented.

- (90) The professor spent hours talking about his trip to Leiden. He loved the ATMOSPHERE
of *the town*.

(adapted from van Deemter 1994)

The final type of givenness is *concept-givenness*. In brief, an expression is *concept-given* if its meaning is entailed by a prior expression (modulo existential closure). The most trivial case is when the expression is identical to an antecedent. Deaccenting is licensed in such cases even if the antecedent and target do not co-refer (or do not refer at all), as (91) demonstrates.

- (91) A: If Susan owns a car, she must be rich.

B: Well anyway, you don't NEED *a car* in New York City.

(van Deemter, 1994)

However, concept-givenness can also give rise to deaccented constituents that do not have identical correlates in the antecedent. One case cited by van Deemter is when the target is a hypernym of an expression in the antecedent. For instance, *string instruments* in (92) is deaccented because it is a hypernym of *viola*.

(92) Bach wrote many pieces for viola. He must have LOVED *string instruments*.

(van Deemter, 1999)

The reasoning behind this example is that Bach's having written pieces for viola entails that he wrote pieces for string instruments, so *string instruments* is (concept-)given and deaccentable.

Conversely, hyponyms are generally not deaccentable because they are not entailed by their antecedents. For instance, in (93a), *the viola* cannot be deaccented because Bach's writing pieces for string instruments does not entail that he wrote pieces for viola. Instead, the sentence must be realized as (93b).

(93) a. Bach wrote many pieces for string instruments. #He must have LOVED *the viola*.

b. Bach wrote many pieces for string instruments. He must have loved the VIOLA.

(van Deemter, 1999)

van Deemter proposes that the three types of givenness are equivalent in terms of their effect on a sentence's focus configuration and, by extension, its pronunciation. However, there are differences in interlocutors' ability to remember different types of givenness that can influence whether expressions are actually marked as given and deaccented. Lexical givenness, of course, is always marked. Between concept- and object-givenness, however, van Deemter proposes that concept-givenness is subject to stricter "memory limitations" than object-givenness. In effect, words and expressions are more transient than reference relations.

Baumann and Riester (2012) expand on this distinction in their givenness annotation scheme, proposing that once a referent is evoked, any subsequent co-referential expression in the same discourse is “referentially given,” while identical expressions can become “lexically new” again within the same discourse if they are separated by a sufficient number of structural units.

Schwarzschild: GIVENness and AVOIDF

Schwarzschild (1999) presents an account of accent licensing that aims to improve on the Selkirk-style model in terms of parsimony. Selkirk’s system invokes a two-part mechanism for focus interpretation, where non-maximal F-marked nodes must be non-given, and F-marked nodes that are not dominated by an F-marked node are interpreted as (contrastive) focus. Schwarzschild uses a new notion, GIVENness, to unify the licensing conditions for maximal and non-maximal F-marking, namely that every constituent must be GIVEN.

GIVENness is defined in terms of the *existential F-closure* of a constituent. Similar to the alternative semantics used by Rooth, $\exists$ F-Closure informally involves replacing any focused elements in the constituent of interest with a variable of the same type. In this case, the variable is existentially bound, so the resulting meaning is a proposition. For the configuration $\text{KIM}_F \text{ won}$, the $\exists$ F-Closure is $\exists x[x \text{ won}]$, or the proposition that someone one. For the configuration Kim WON_F , the $\exists$ F-Closure is $\exists Q[\text{Kim } Q]$, or the proposition that Kim did something (has some property).

When the constituent of interest is not a clause, $\exists$ F-Closure also introduces additional existentially bound variables in order to form a proposition. So, for the configuration *won the RACE*_F, the focused element is replaced with a variable, e.g., *won x*. Then, another variable is introduced, and both variables existentially bound, so the resulting meaning is a proposition, i.e., $\exists x \exists y[y \text{ won } x]$.

A constituent is GIVEN if the context supplies a salient meaning that is in the focus semantic value of the constituent in question (if the constituent denotes an individual), or the context supplies a salient meaning that entails the $\exists$ F-Closure of the constituent. As men-

tioned above, the principal requirement in Schwarzschild’s system is that every constituent must be GIVEN; that is, any F-marking, as well as the absence of F-marking, must be licensed by the prior context in terms of GIVENness.

This is straightforward for cases of overt repetition. Consider (94).

(94) Christina embraced Laura, and RON *embraced Laura*.

First, consider the terminal nodes in the second clause. RON_F is GIVEN because it denotes an individual, and an element of its focus semantic value (an individual, e.g., *Christina*) is supplied in the context. *embraced* is GIVEN because its $\exists F$ -Closure, $\exists x \exists y [x \text{ embraced } y]$, is entailed by the first clause, *Christina embraced Laura*. *Laura* has the focus semantic value $[Laura]$, and *Laura* is instantiated in the first clause.

Turning to the non-terminal nodes of the second clause, the $\exists F$ -Closure of *embraced Laura* is $\exists x [x \text{ embraced Laura}]$, which is entailed by *Christina embraced Laura*, so it is GIVEN. Finally, $\text{RON}_F \text{ embraced Laura}$ has the $\exists F$ -Closure $\exists x [x \text{ embraced Laura}]$, which is entailed by *Christina embraced Laura*, so the full clause is GIVEN.

Schwarzschild’s system can also deal with some more complicated relations between givenness, focus, and emphasis. Consider (95).

(95) A: Who did John’s mother recommend?

B: She recommended JOHN.

The intuitive question here is why JOHN is prosodically emphasized when it is given in the preceding question. Consider the alternative utterance, *She recommended John*, where *John* is deaccented. At the level of the terminal node, this configuration should be licensed – the antecedent supplies a meaning, *John*, that is in the focus semantic value of the deaccented *John*.

The explanation for the unavailability of the alternative utterance stems from Schwarzschild’s requirement that every constituent in a particular focus configuration be GIVEN. This requirement is violated by the higher nodes in the alternative utterance. Consider the VP

recommended John, where neither constituent is F-marked. This constituent has an $\exists$ F-Closure of $\exists x[x \text{ recommended John}]$, which is not entailed by anything in the discourse context. By contrast, the felicitous constituent *recommended JOHN_F* has the $\exists$ F-Closure $\exists x\exists y[x \text{ recommended } y]$, which is entailed by the antecedent given an appropriate semantics for questions. Thus, *recommended John* is not GIVEN, while *recommended JOHN_F* is, and only the second configuration is licit.

The other major component of Schwarzschild’s model, which is not of critical theoretical importance here, is his means of accounting for the generalization that given elements generally are not F-marked (barring atypical configurations like (95) above). Where Selkirk introduced an overt grammatical requirement that given elements cannot be F-marked, Schwarzschild does not include any such rule – indeed, such a rule intuitively would block configurations like (95).

As an alternative, Schwarzschild introduces a violable syntactic constraint, AVOIDF. This is a markedness constraint that penalizes F-marking. AVOIDF is prioritized below the principle that every constituent be GIVEN. The result is that elements are only F-marked in order to satisfy GIVENness. In the typical case, an element is F-marked because it is non-given, as RON is in (94) above. This F-marking is necessary, informally, because non-given elements must be replaced with variables in order for an antecedent with a different element in that position to entail the relevant $\exists$ F-Closure.

As a general rule, given elements will not be F-marked, because such configurations are penalized by AVOIDF. However, given elements can be F-marked if it is necessary to satisfy GIVENness. This is demonstrated in (95) above, where given *John* must be F-marked in the response because failing to do so would make the non-terminal nodes in the response non-GIVEN.

Notable for the current dissertation is Schwarzschild’s direct inclusion of an entailment relation in the definition of GIVENness. As for Selkirk’s approach, this allows for cases of nonidentity between a deaccented element and an antecedent where, for instance, the

deaccented element is a hypernym of the antecedent, as in (96).

- (96) Bach wrote many pieces for the viola. He must have LOVED *string instruments*.
(van Deemter, 1999)

The deaccenting of *string instruments* would not be licensed in a theory requiring exact identity between a deaccented element and an antecedent, since *string instrument* is not instantiated in the antecedent. However, such an example is ruled in under Schwarzschild's theory, since the antecedent element *violin* entails the $\exists F$ -Closure of *string instruments*.

Schwarzschild does note that the entailment relation is meant to be relatively broad. For instance, he notes (97).

- (97) A: John got the job.
B: I know. They WANTED *a New Yorker*.
(Schwarzschild, 1999)

In this case, the deaccenting of *a New Yorker* clearly relies on knowledge shared by the speaker and hearer that John is a New Yorker. For this reason, the entailment relation is construed as holding between a discourse context, not just the linguistic antecedent, and the $\exists F$ -Closure of the relevant focus consideration. In other words, deaccenting can be licensed when the constituent of interest is entailed by a linguistic antecedent, but also presumably by information that is salient in the broader discourse context or even information in the discourse participants' common ground.

Beaver and Clark: From GIVENness to activation

Beaver and Clark (2008) do not focus extensively on anaphoric deaccenting, but do discuss Schwarzschild's approach briefly. They note some problematic examples where material that is certainly GIVEN is accented, as in (98).

- (98) A: Sandy fed Fido this morning.
B: Um, wait, she DIDN'T *feed* FIDO, *she fed* BUTCH.
(Beaver and Clark, 2008)

In the clause *she didn't feed Fido*, Schwarzschild's account predicts only the accent on DIDN'T, since *Fido*, *feed*, and *feed Fido* are GIVEN. Schwarzschild's explanation for accented GIVEN material is to propose that it is a result of mandatory accentuation in FOC-marked phrases that do not otherwise contain an accent. However, in (98), the phrase does contain an accent on DIDN'T, and FIDO is accented all the same.

Beaver and Clark propose to replace GIVENness with a notion they refer to as *activation*. Constituents must be F-marked when they are insufficiently activated. A constituent is insufficiently activated either if it is not GIVEN, or if the speaker wishes to evoke alternatives for it. The proposal is thus functionally equivalent to Schwarzschild's model, except that constituents like FIDO in (98) are correctly predicted to be accented.

The system of activation is not spelled out further than this sketch. Beaver and Clark do note that there are potentially several complicating factors, noting as an example that visual salience in fully realized discourse contexts might affect what counts as sufficiently activated. The terminology of *activation* also recalls Chafe's (1974; 1994) approach to givenness, which will be discussed in more detail below. In this approach, information counts as given (and deaccentable) essentially when the speaker assumes that it is currently active in the mind of the hearer. Thus, the concept of activation somewhat blurs the line between defining givenness (or GIVENness) in purely structural terms and allowing less formal, context-driven assessments of givenness to play a role in accent assignment.

Baumann and Riester: RefLex

Baumann and Riester (2012) argue for a highly articulated system for annotating givenness, accessibility, and novelty. The defining feature of their annotation system, called RefLex, is that it distinguishes between two types of givenness: referential givenness, or r-givenness, and lexical givenness, or l-givenness. They note that treating these two types of givenness separately dates at least to van Deemter's (1994; 1999) notions of object-givenness and concept-givenness. They also argue that Schwarzschild (1999) implicitly made such a distinction by providing separate definitions of GIVENness for type *e* utterances that is based

exclusively on co-reference with an antecedent.

The critical observation behind RefLex is that r-givenness and l-givenness are both sufficient conditions for deaccenting, but neither is necessary. This is demonstrated in (99) and (100).

(99) A: Did you see Dr. Cremer to get your root canal?

B: Don't remind me. I'd like to STRANGLE *the butcher*.

(Büring, 2007)

(100) A: Why do you study Italian?

B: I'm MARRIED *to an Italian*.

(Büring, 2007)

In (99), *the butcher* can be deaccented based solely on its intended co-reference with *Dr. Cremer*, despite the fact that the meaning of *butcher* is not available in the discourse. Conversely, in (100), *Italian* in B's reply does not co-refer with an entity evoked in the prior discourse, but it is deaccentable by virtue of its lexical connection in meaning to the instance of *Italian* denoting the language in A's question.

A critical observation that arises from these examples is that, at least from the perspective of deaccenting, givenness in the referential sense and givenness in the lexical sense can be orthogonal to one another. As a result, Baumann and Riester propose separate annotation schemes for r-givenness and l-givenness.

Both schemes are highly detailed, as Baumann and Riester hold that a number of different factors influence the givenness of an item in the discourse. Of particular interest here are annotations marking an item as given, since this is taken as sufficient for deaccenting. At the referential level, the most straightforward givenness annotation is labeled *r-given* and is assigned when an expression co-refers with an antecedent in the previous discourse. In (101), *the man* in the second sentence is r-given.

(101) I met a man yesterday. The man told me a story.

(Baumann and Riester, 2012)

Beyond this, there are two other given annotations. *r-given-sit* is assigned to a constituent whose referent is present in the external context that are available in any discourse by default. This is assigned primarily to pure indexicals like *you*, *I*, *tomorrow*, and *here*. The label of *r-given-displaced* is assigned to a co-referring expression whose antecedent appeared more than five intonational phrases or clauses prior in the discourse. It is not clear that *r-given-displaced* constituents are analyzed as deaccentable, but Baumann and Riester suggest that once mentioned, a constituent cannot become fully *r-new* again for the rest of the discourse.

Two additional annotations that are potentially relevant to prosody are *r-environment* and *r-bridging*. *r-environment* is assigned to a constituent that refers to an item in the text-external context that is specific to a particular discourse, and generally depends on gestural deixis. The constituent *This chair* in (102) is labeled *r-environment*.

(102) This chair [pointing] is wobbly.

(Baumann and Riester, 2012)

r-bridging is assigned to constituents that are not co-referential, but the existence of whose referents is easily inferred from the prior scenario. In (103), *the referee* is *r-bridging* because football matches generally have referees.

(103) Recently, at the football match, the referee was wearing orange socks.

Baumann and Riester do not propose that *r-environment* and *r-bridging* constituents are deaccentable. However, it is worth questioning whether they may actually be deaccentable under certain conditions. For instance, Rochemont's (1986) hospital example, repeated as (104), relied on the listener's knowledge that the speaker routinely goes to a hospital, e.g., as an employee, to license deaccenting of *hospital*. This might best be thought of as a case of deaccenting an *r-bridging* constituent.

(104) I ran into JOHN at the *hospital* this morning.

At the lexical level, RefLex tries to capture the intuition that expressions can be lexically given not only when they are identical to an antecedent, but also entailed by one (modulo existential closure). An expression is *l-given-same* if it is string identical to an expression that was used earlier in the discourse. For instance, *the man* in (101) is l-given-same in addition to being r-given. An expression is not l-given-same if it does not share a sense with its potential antecedent; for instance, *ball* referring to a spherical object is not l-given-same if it appears after *ball* referring to a dance.

An expression is *l-given-syn* if it is synonymous with an antecedent. For example, *lift* would be l-given-syn if used after *elevator*. An expression is *l-given-super*, and therefore deaccentable, if it is lexically superordinate to a previous noun, e.g., as a hypernym. In (105), *string instruments* is l-given-super based on its relationship to *viola* and is deaccentable.

(105) Bach wrote many pieces for viola. He must have loved string instruments.

(van Deemter, 1999)

Conversely, an expression is merely *l-accessible-sub* if it is lexically subordinate to its antecedent, as *viola* is in (106). Constituents annotated as l-accessible-sub are generally not deaccentable.

(106) Bach wrote many pieces for string instruments. He must have loved the viola.

(van Deemter, 1999)

Another potentially relevant annotation at the lexical level is *l-accessible-other*. This is assigned when an expression is lexically related to a prior expression with an unclear hierarchical relationship, e.g., *lawyer* and *courtroom*. This annotation seems to leave room for more world knowledge- or scenario-driven inferences to mark expressions as accessible, such as *call a Republican* and *insult*, although Baumann and Riester do not overtly state that accessible expressions should be deaccentable.

Finally, if an expression has not been used and is not related to an expression that has been used in the previous five intonational phrases or clauses, it is *l-new*. This is in contrast to the referential level, where an expression can never be r-new if its referent has been evoked at any point in the current discourse.

As mentioned above, Baumann and Riester argue that r-givenness and l-givenness are both sufficient conditions to license deaccenting, but neither is necessary. They investigate this claim by annotating two corpora, one of spontaneous speech and one of read speech, using the RefLex scheme and coding the nominal expressions in the corpora for their accent status. Table 5.1 summarizes the results of this investigation in terms of the percentage of nominal expressions that were deaccented for different combinations of r- and l-givenness.

	Spontaneous	Read
Not given	<4%	0%
r-given only	18%	24%
l-given only	31%	10%
r- and l-given	17%	71%

Table 5.1: Percentage deaccenting for r- and l-givenness (Baumann and Riester, 2012)

Although the results of this investigation do not show uniform deaccenting of either r- or l-given constituents, the difference between the prosodic realization of expressions with at least one type of givenness and expressions that are neither r- nor l-given is telling, and Baumann and Riester take the results as preliminary evidence in favor of the RefLex scheme.

One open question for this approach is how it generalizes beyond nominal expressions. Baumann and Riester make it clear that the scheme is only designed to handle nominals, yet it is known that deaccenting can target other types of constituents, like verbs. One possibility is that other constituents do not participate in the referential annotation scheme at all, since they cannot refer, but that the lexical level works similarly to nominals. For instance, *embraced* might be l-given-super after *hugged* if embracing acts are a superset of hugging acts. Likewise, as mentioned above, *insult* might be l-accessible-other after *call a Republican*. Determining how to extend the scheme beyond nominal expressions is necessary

to provide a complete account of deaccenting, and these questions will be important in the next two chapters, which critically target deaccenting of verbs, rather than nouns, under identity and nonidentity.

Büring: F-marking, G-marking, and clarifying the role of context

Büring (2016) reviews Rooth’s (1985; 1992a) alternative semantics, Selkirk’s (1995) focus projection, and Schwarzschild’s (1999) GIVENness approaches to emphasis licensing. Büring argues that Schwarzschild’s GIVENness relation is compatible with other focus- and givenness-marking schemes (besides the “stacked F-marker” approach used by Selkirk and Schwarzschild). He proposes several changes to the stacked F-marker approach that he argues can capture the data equally well while interfacing better with other areas of linguistic theory.

First, Büring proposes a system that does not use stacked F-markers to account for the phenomenon of focus projection. Recall that Selkirk proposed that an accented word must host an F-marker, but that F-marking of internal arguments also licensed F-marking of heads, and F-marking of heads licensed F-marking of the corresponding phrase. The “focus” of the sentence, in the traditional sense, is FOC, an F-marked constituent that is not itself dominated by an F-mark.

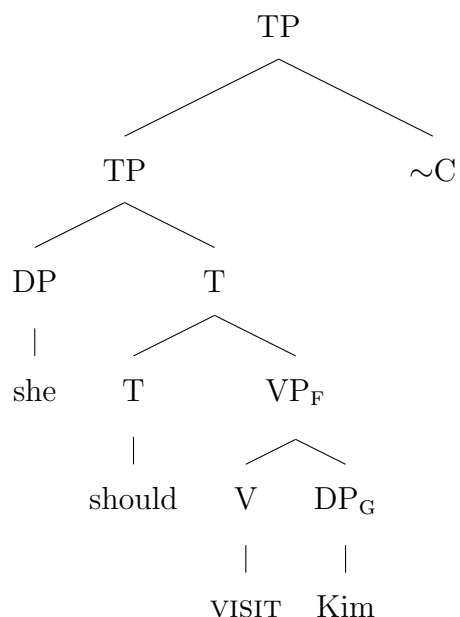
By contrast, Büring’s system uses one F-mark to denote a single focus. As a result, the phenomenon of focus projection must be explained by some means other than the projection of the F-feature as outlined by Selkirk, as will be discussed below. He argues that such a system allows for a relatively looser relation between F-marking and prosodic realization, with principles of “default prosody” responsible for the bulk of the work of determining accent placement in a sentence. Further, he notes that the convention in the majority of work on information structure is to use a single F-marker to indicate the focus of a sentence, and proposes that a system using single F-markers interfaces with this literature more readily.

In order to more narrowly tie the F-marker to the notion of “focus,” Büring proposes to use a separate marker, G, to denote discourse-given material, as shown in (107a). This

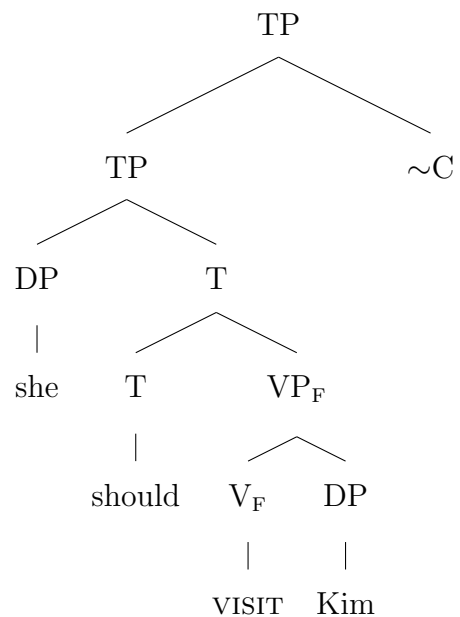
contrasts with, for instance, Selkirk’s system, which indicates givenness by the absence of an F-mark on a constituent that lies within a FOC, as shown in (107b).

(107) What should Kim’s mother do?

a.



b.



(Büring, 2016)

Büring proposes different conditions on the use of F- and G-markers. For an F-marker to be felicitous, the argument of $\sim$ – that is, the “Contextually Supplied Meaning” – must

make the sister of $\sim$ – the constituent containing the focus – GIVEN. That is, the structure shown in (108) presupposes that C makes δ GIVEN, i.e., that $\exists\text{Clo}(\text{C})$ entails $\exists\text{F-Clo}(\delta)$.



By contrast, the G-marker merely presupposes that the marked constituent is given (not GIVEN) – that is, that there is some semantic antecedent for the constituent available in the discourse, as assumed by Rochemont (1986).

Büring introduces a new set of conditions governing the use of F- and G-markers in order to enforce their presence in the appropriate domains. These conditions are shown in (109) through (111).

(109) MAXIMIZE BACKGROUND

In any tree, maximize the number of (non-synonymous) constituents that are in the background.

(110) BACKGROUND in an F+G-marking system

A constituent E is IN THE BACKGROUND in a tree T iff E is c-commanded by a $\sim\text{C}$ in T , and not (part of) a focus of that $\sim\text{C}$, or if E bears, or is dominated by, a G-marker in T .

(111) FOCUS ON AN F-DOMAIN

An F-marked constituent E is a FOCUS of a $\sim\text{C}$ in tree T if $\sim\text{C}$ minimally c-commands E in T .

(Büring, 2016)

Büring argues that this system achieves multiple desirable results. First, it addresses what Büring calls the problem of over-focusing. Recall from the discussion of the focus projection

framework that a clause with a particular prosodic configuration can have an ambiguous focal interpretation – for instance, *Mary bought a book about BATS* can be interpreted as having focus on the argument of the preposition, the PP, the direct object, the VP, or the full sentence. Büring contends that an appropriate theory of F-marking will explain why listeners do not infer larger foci than are licensed in a particular discourse environment. For instance, following the question *What did Mary buy a book about?*, although there is no prosodic incompatibility with interpreting the reply with VP focus, Büring suggests that the F-marking mechanism should enforce selection of focus on the prepositional argument.

Büring’s system accomplishes this using the principle of MAXIMIZE BACKGROUND. Non-F-marked material that is c-commanded by $\sim C$ is in the background. Thus, it is generally most desirable to parse a structure with a maximally large focus domain (i.e., with $\sim C$ high in the tree) and small foci, since constituents above the focus are not F-marked but will still be c-commanded by $\sim C$. Selection of a larger focus will only occur when it is motivated by independent discourse considerations (i.e., as the appropriate answer to a *wh*-question, via a constraint like MAXIMIZE PRESUPPOSITION).

Second, the system enforces G-marking of given constituents, which results in their not being prosodically emphasized as long as there is another constituent within the focus (dominated by F) that can bear prosodic prominence. This is because a constituent counts as being in the background as long as it is G-marked. Thus, non-emphasis of given constituents is mandatory unless it prevents a pitch accent from surfacing within the focus; whether the generalization that all such given constituents are de-emphasized will be a topic of investigation in the experiments presented below.

Turning to the question of de-emphasis under nonidentity, there are two types of backgrounded, de-emphasized constituents in Büring’s system: those dominated by an F-mark but not receiving an accent by the rules of prosody, and those bearing or dominated by a G-mark. Büring is clear that F-marking is licensed by GIVENNESS, whereas G-marking is licensed by mere givenness. Of interest here are G-marked constituents; it does not make sense

to discuss constituents dominated by an F-mark as being non-identical to an antecedent, as the notion of antecedence relevant for F-marking, GIVENness, involves replacing F-marked constituents with variables.

The requirement for a G-marked constituent is mere givenness. By this, Büring means roughly the “semantic antecedence” notion adopted by Rochemont (1986), namely that there be a contextually salient meaning whose existential closure entails the existential closure of the target constituent’s denotation. He explicitly lists synonyms, hypernyms, and coreferential expressions as constituents whose de-emphasis can be licensed by the presence of a non-string-identical antecedent. Thus, deaccenting is licensed directly in the grammar for any given constituent, where givenness minimally covers identical and entailed material.

Büring seeks to clarify how, in a model like Schwarzschild’s, world knowledge can be modeled in terms of privative givenness. In examples like (97), repeated in (112), *New Yorker* is deaccented because John is a New Yorker, and the speaker assumes that mentioning John makes the concept of a New Yorker sufficiently salient.

(112) A: John got the job.

B: I know. They WANTED a *New Yorker*.

The puzzle is how to represent such information in the definition of GIVENness. Büring reviews several attempts that would not suffice, among which: “An expression E is given if there is a contextually salient meaning A so that *shared assumptions of the participants* together with the $\exists\text{Clo}$ of A entail the $\exists\text{Clo}$ of E.” The reason this is inadequate is because it would allow any information shared by the speaker and hearer to be treated as given and deaccented, as in (113).

(113) A: You should come over for dinner next week!

B: #Sorry, I’m GOING to *London* next week.

(Büring, 2016)

Here, *London* is deaccented based on the shared knowledge that London exists, which is incorrectly predicted to be felicitous according to the sketch of GIVENness given above. Since (113) relies entirely on background assumptions, a second attempt at reformulating GIVENness proposes that the shared assumptions alone cannot qualify material as GIVEN. However, Büring notes that this disqualifies (112), where just the fact that John is a New Yorker entails that there is a New Yorker.

Büring proposes that the role of contextual knowledge should instead be defined at the level of the F(ocus)-domain: “An F-domain is licensed if there is a contextually salient meaning A so that shared assumptions of the participants together with the $\exists\text{Clo}$ of A, but not [the shared assumptions] alone, entail the $\exists\text{F-Clo}$ of the F-domain.” Incorporating contextual knowledge at this level enforces that the knowledge reflects critical relationships between constituents, e.g., *there was some relation they had to a New Yorker* rather than *there is a New Yorker* in (112), or *B has some relationship to London* rather than *London exists* in (113).

5.2.2 *The accommodation approach*

The second broad approach to deaccenting licensing is distinct from the grammatical approach in that it appeals directly to lexical identity, rather than givenness and inferability, in the grammar of deaccenting. Each of the models below holds that the grammar of deaccenting requires identity between the target constituent and a linguistic antecedent. Deaccenting of nonidentical material can be acceptable, but depends on an extragrammatical operation. The core proposal is that listeners who encounter ungrammatically deaccented material will construct a covert alternative antecedent containing a correlate identical to the target; that is, the covert antecedent is one that would felicitously license the observed deaccenting under identity. Presumably, this process is pragmatically mediated, with construction of a covert antecedent occurring only if the content of this structure is considered reasonable in the discourse.

Tancredi: Pragmatic incrementation

In developing his account of emphasis licensing, Tancredi (1992) initially considers a notion of focus alternatives (“focus-related topic”) similar to that familiar from Rooth (1985, 1992b) – that is, the set of structures generated by replacing all focused nodes in the domain under consideration with variables of the same type. However, examples like (114) lead him to propose a different analysis.

(114) Mary was crying, so JOHN COMFORTED *Mary*.

(Tancredi, 1992)

Assuming *John* and *comforted* are marked as separate foci, Tancredi argues that the “replace with variables” strategy would require an antecedent of the form *R Mary*, where the preceding two constituents have been combined into a one-place property with *Mary* as an argument. He points out that this would be a violation of the principle of strict compositionality.

Instead, Tancredi suggests that the “focus-related topic” of a focus configuration should be constructed from those constituents in the configuration that are deaccented (that is, not F-marked). Consider (115).

(115) *John* BELIEVES *that Mary’s alive*.

(Tancredi, 1992)

Tancredi argues that this sentence is felicitous in a discourse with prior instances of *John* and *Mary is alive*, as shown in (116), but not one that merely has separate instances of *John*, *Mary*, and *is alive*, as in (117).

(116) John doesn’t know if Mary is alive. However, *John* BELIEVES *that Mary’s alive*.

(Tancredi, 1992)

(117) Mary doesn’t know if John is alive. #However, *John* BELIEVES *that Mary’s alive*.

(Tancredi, 1992)

Tancredi notes that this fact follows straightforwardly from the “replace with variables” strategy, which would require an antecedent of the form *John R Mary is alive*. However, given the separate compositionality concerns, he proposes the following mechanism for calculating the felicitousness of a focus configuration. First, substitute a variable for any focused constituent – since only *believes* is focused here, this yields *John R Mary is alive*. Next, identify the *Complete Functional Complexes* formed by the remaining elements. In this case, this is the set consisting of the NP *John*, the NP *Mary*, and the clause *Mary is alive*. Finally, determine the smallest subset of this set that includes all of the non-variable elements, which in this case is the set consisting of the NP *John* and the clause *Mary is alive*.

By Tancredi’s formulation, a focus configuration is felicitous if every element of the focus-related topic (e.g., the set of *John* and *Mary is alive* in the current example) is “instantiated” in the active context. Informally, this means that felicitousness is predicated in terms of deaccented constituents. Any constituent that is deaccented in the configuration under consideration must be overtly instantiated in the antecedent. However, the “smallest subset” step of the process of determining a focus-related topic enforces that only the largest possible deaccented constituents are checked. That is, in (117) above, it does not suffice that *Mary* and *is alive* are separately instantiated in the antecedent, because the “smallest subset” step requires that the constituent *Mary is alive* is instantiated in the discourse context.

Of particular interest to the current work is the notion of *instantiation*, which Tancredi discusses in some detail. Similar to other authors, Tancredi considers examples where a deaccented constituent does not have an identical correlate in the antecedent, but rather is inferable from part of the antecedent, as in (118).

(118) Yesterday, Sam killed Bill. Today, JOHN *died*.

(Tancredi, 1992)

In (118), *died* is deaccented even though it is not overtly instantiated in the antecedent. Informally, it is clear that the felicitousness of this example is driven by the implicit inference that *Bill died* made available in the antecedent clause. Tancredi questions whether to account

for this type of utterance by treating the inferable element as “instantiated” in the context, or by requiring exact identity for “instantiation” and deriving the felicitousness through another process.

Ultimately, Tancredi argues that instantiation in the context should require exact identity, and that deaccenting licensed via inferability is achieved through “pragmatic incrementation” of the linguistic context. To argue for this approach, he cites examples where trivially inferable elements are deaccented. For instance, since any sentence entails a tautology, *Sam killed Bill* entails *Either Sam is exactly six feet tall or Sam is not exactly six feet tall*. However, the utterance in (119) is not felicitous with the indicated prosody, even though a constituent of the form *is exactly six feet tall* can be inferred from the linguistic context.

(119) Yesterday, Sam killed Bill. And now, JOHN’s *exactly six feet tall*.

(Tancredi, 1992)

Tancredi then considers the possibility of restricting the list of instantiated inferable constituents to include only “logical” entailments of the antecedent, but continues to cite (120) as a counterexample.

(120) John called Mary a Republican. Then, BILL *insulted Mary*.

(Lakoff, 1968; Tancredi, 1992)

As discussed above, the challenge presented by this utterance is that it is driven by an inference that is clearly not a logical entailment of the antecedent. Rather, individual speakers and hearers will differ from one another regarding whether they agree that *insult x* can be inferred from *call x a Republican*, and the felicitousness of (120) relies on a determination that this inference is valid.

Thus, the licensing of deaccenting via inferability needs to be restricted enough that trivially inferable constituents like *is exactly six feet tall* in (119) are ruled out, but constituents that are not strictly entailed, like *insulted* in (120), can be ruled in. Tancredi argues that it is impossible to find a semantically meaningful notion of inferability meeting these criteria to

incorporate into the definition of instantiation in the discourse context, and suggests instead that the process of deaccenting via inference must be pragmatically driven.

Tancredi characterizes his analysis in terms of conversational implicature – meanings that are made available in the discourse context via an implicature can be deaccented in a particular focus configuration. It may not be correct to characterize the process in this way – for instance, it is not clear that the inference *Bill died* from *Sam killed Bill* is an implicature – but the pragmatic mechanism Tancredi proposes is nevertheless clear.

Uttering a sentence with a deaccented constituent is felicitous only when that constituent is instantiated in the discourse context, which here means that it appears overtly in the recent linguistic context. Thus, utterances like (118) and (120) should strictly be infelicitous (or, perhaps more appropriately, ungrammatical), because they contain deaccented constituents that do not appear in the prior linguistic context.

However, hearers assume that speakers are observing the Cooperative Principle (Grice, 1975). As such, they should be invested in clear communication, and in general should avoid infelicitous or ungrammatical utterances. This principle triggers a search for an augmented discourse context in which the deaccented constituent can be construed as being instantiated; Tancredi indicates that this augmentation might be thought of in terms of accommodation (Lewis, 1979). Thus, for (118), the listener might accommodate a new antecedent that includes the constituent *Bill died*, and likewise *John insulted Mary* for (120). While the deaccented constituents in each sentence are not strictly instantiated in the overt linguistic antecedent, they are instantiated in the accommodated antecedent, which Tancredi argues is sufficient for marking the utterance as felicitous (or rather, acceptable).

Crucially, this approach relies on the principles of rational communication to determine which inferences “count” as eligible for deaccenting. Evidently, the inferences that *Bill died* and *John insulted Mary* are sufficiently available from the antecedent clauses in (118) and (120) that antecedents instantiating them can be accommodated, while an antecedent instantiating *Either Sam is exactly six feet tall or Sam is not exactly six feet tall* cannot be

accommodated in the case of (119). Tancredi argues that this follows from general principles of rational communication. It is reasonable to assume that in the case of the former two instances, the speaker would expect the hearer to draw these inferences as a result of their utterance. By contrast, while the inference in the case of (119) is logically licensed by the antecedent, it is not relevant to the antecedent utterance. Thus, the speaker should not expect that the hearer has accessed this inference upon hearing the antecedent, and the hearer, knowing this, is likely not prepared to encounter a focus configuration that requires instantiation of the inferred material in an accommodated antecedent.

Finally, it is worth noting that beyond these cases of relatively “minor” nonidentity, where a deaccented constituent can be inferred from an antecedent constituent, Tancredi also notes cases of “radical” nonidentity, where a constituent is deaccented even in the complete absence of a linguistic antecedent. Consider (121).

(121) I HATE *computing over telephone lines*.

(Tancredi, 1992)

Tancredi proposes that this sentence may be felicitous with the indicated prosody even in certain discourse-initial contexts – that is, with no prior linguistic material to use as an antecedent (or from which to accommodate an antecedent). He uses this example to argue in favor of his proposal of context augmentation. Since there is no prior linguistic material in this discourse, it seems straightforward that deaccenting of inferable material is not licensed by a grammatical operation that directly considers inferences made available by the antecedent, but rather by a pragmatic operation allowing hearers to posit new linguistic structures against which identity can be calculated. Tancredi does not discuss such discourse-initial examples in detail, and it is worth noting here that there seems to be room for future empirical investigation of discourse-initial deaccenting, including the conditions on its use and the interaction between the salience of information in the nonlinguistic context and the felicitousness of deaccenting relying on that information.

Fox: Parallelism and accommodation-seeking material

Fox (2000), similar to Rooth (1992a), seeks to delineate the similarities and differences between the identity relations that are in play for the licensing of ellipsis and of deaccenting. In contrast to Rooth, though, Fox suggests that the identity relations licensing ellipsis and deaccenting are actually the same, and that the flexibility of deaccenting under nonidentity compared to ellipsis is actually driven by the relative ease of accommodating new antecedents for deaccented constructions compared to elliptical ones.

Fox notes that in addition to canonical cases of deaccenting under exact identity with an antecedent, deaccenting is occasionally licensed in the presence of a related antecedent that is not identical to the deaccented material. However, deaccenting under nonidentity is subject to an asymmetry, as demonstrated in (122a) and (122b).

- (122) a. First, Bill called Mary an idiot. Then, JOHN *insulted her*.
b. *First, John insulted Mary. Then, BILL *called her an idiot*.
(Fox, 2000)

(122a) is by now a familiar case of deaccenting under nonidentity. In the second sentence, the deaccented *insult* intuitively indicates that an instance of *insult* should be available in the discourse. It is easy for a rational hearer to infer from the first sentence, *Bill called Mary an idiot*, that Bill also insulted Mary – there are relatively few discourses in which this occurs where it would not constitute an insult.

By contrast, (122b) is not felicitous. The deaccented *called her an idiot* in the second clause indicates that some other instance of *call an idiot* should be available in the discourse. However, the only available antecedent is *insult*. *call an idiot* cannot reasonably be inferred from *insult*, since there are many ways to insult someone without calling them an idiot. On the basis of this asymmetry, Fox proposes that one mechanism for licensing deaccenting relies on an entailment (or other inference) relation between the antecedent and the constituent of interest. Deaccenting is licensed when the constituent in question can be inferred from a

constituent in the antecedent, but not when the antecedent is inferable from the deaccented constituent, as in (122b).

Fox proposes a two-pronged account of emphasis (and ellipsis) licensing, where deaccenting can be licensed either via *direct parallelism* or *indirect parallelism* with an antecedent. According to *direct parallelism*, a constituent can be phonologically reduced (deaccented) if there is a constituent in the linguistic antecedent that is an “appropriate alternative” to the constituent of interest – that is, if there is a linguistic antecedent that is in the focus semantic value of the constituent of interest.

According to *indirect parallelism*, deaccenting is licensed when the linguistic antecedent, together with certain presuppositions shared by the speaker and hearer, entails a focus alternative of the constituent of interest. Thus, comparable to the other accounts reviewed so far, Fox derives the felicitousness of (122a) using an inferencing relation linking *call an idiot* and *insult*. Bill’s calling Mary an idiot doesn’t logically entail that he insulted her, but in most contexts, this inference will be considered reasonable by both speaker and hearer. Thus, the antecedent, together with independent world knowledge on the part of the discourse participants, “entails” (informally) that Bill insulted Mary. *Bill insulted Mary* is a focus alternative of *JOHN insulted Mary*, so the focus configuration in the second sentence of (122a) is licensed.

Critically, Fox operationalizes indirect parallelism in terms of accommodation. Deaccenting of an inferable constituent is not directly licensed by the entailment relation (as in, for example, Rooth’s proposal to include inferencing relations in the semantics of $\sim$). Rather, Fox explicitly refers to the focus alternative that must be entailed by the antecedent and participants’ world knowledge as an accommodated antecedent, and in order for deaccenting to be licensed by indirect parallelism, the relation between the accommodated antecedent and the deaccented constituent must satisfy the requirements of direct parallelism. Thus, there is a linguistic object that must be constructed, and against which exact identity with the deaccented constituent must be calculated, in order for deaccenting via inference to be

licensed.

Two key components constrain Fox's mechanism of antecedent accommodation. The first is the notion of *accommodation-seeking material*. One of Fox's goals is to explain why it is relatively easy to license deaccenting under nonidentity, as in (122a), but not ellipsis, as in (123).

(123) First, Bill called Mary an idiot. Then, JOHN *did*.

(Fox, 2000)

In (123), the only interpretation of the ellipsis site is that John called Mary an idiot, not that John insulted her (that is, the inferable content from (122a)). Fox's indirect parallelism mandates that in order for nonidentity between an antecedent and either a deaccenting or ellipsis site to be licensed, there must be some inferable accommodated antecedent such that direct parallelism is satisfied. However, in Fox's system, accommodation must have a trigger. He introduces the notion of *accommodation-seeking material*, which is any non-F-marked (i.e., deaccented) material that was not present in the overt antecedent.

When material is not F-marked in anaphoric clauses, it signals to the listener that the non-F-marked material is retrievable from the discourse. In Fox's system, this should strictly mean that an identical constituent is available in the linguistic antecedent. When this is not the case, hearers, knowing that the speaker likely intended to produce an acceptable utterance, will look for a reasonable accommodated antecedent such that the non-F-marked constituent is instantiated. Crucially, non-F-marked material is common in deaccenting, but very rare in ellipsis (because any non-F-marked material in ellipsis is by definition deleted). Thus, Fox proposes that the identity conditions on deaccenting and ellipsis are actually the same, and that the relative flexibility of identity for deaccenting compared to ellipsis is driven by the frequent presence of accommodation-seeking material in such constructions.

The second constraint of Fox's accommodation system is the notion of *minimal accommodation*. Essentially, this constraint mandates that the accommodated antecedent must be the simplest antecedent that satisfies the requirements of direct parallelism. A particular

accommodated antecedent *A* cannot be considered if there is another potential antecedent *B*, also satisfying direct parallelism, whose accommodated material is a proper subset of the accommodated material in *A*. In essence, this constraint bans accommodation of antecedents that are more distant than absolutely necessary from the overt linguistic antecedent. This explains why the “insult” reading for (123) cannot be optionally licensed by indirect parallelism, since *insult* can be inferred from *call an idiot*. The overt antecedent already satisfies the requirements of direct parallelism, and the accommodated antecedent *Bill insulted Mary* contains more accommodated material than the overt antecedent. Thus, accommodating *Bill insulted Mary* would not be minimal, and the reading with *insult* in the ellipsis site is not available.

Wagner: Local exclusive alternatives

Wagner (2012) presents a theory of emphasis licensing that depends on the presence of “proper” local alternatives. The critical data points motivating this analysis are examples like (124) and (125).

(124) A: Last week someone shot Smith, the manager of the department store. Today he was finally released from the hospital.

B: So, are things back to normal now?

A: #No, you won’t believe it. Someone SHOT *Smith*. They had to close again!

A: No, you won’t believe it. Someone shot SMITH. They had to close again!

(Wagner, 2012)

(125) A: Last week class ended early because a physics student pulled the fire alarm in the building. What was the reason this week?

B: #You won’t believe it. A STUDENT *pulled the fire alarm*.

(Wagner, 2012)

Wagner argues that these examples demonstrate that prior theories of emphasis licensing (e.g., Schwarzschild, 1999) impose too weak of conditions on the antecedent of a focus

configuration. Consider Schwarzschild’s approach for the reply in (124) where the verb is prosodically emphasized. Schwarzschild’s approach requires that the $\exists$ F-Closure of *someone* SHOT *Smith* (among other constituents) must be entailed by the antecedent. The $\exists$ F-Closure of this clause is $\exists R[\text{someone } R \text{ Smith}]$, which is indeed entailed by the antecedent clause *someone shot Smith*. However, Wagner claims that it is not felicitous to shift prominence from the default position (on *Smith* in the reply) to the verb.

Informally, Wagner argues that the reason stress shifting away from the default (final) position in (124) and (125) is infelicitous is because the emphasized constituents are not “proper” alternatives to the corresponding antecedent constituents. In (124), *someone shot Smith* is not an alternative to *someone shot Smith* in the sense that it does not contrast with the antecedent in any meaningful way. Likewise, STUDENT in (125) does not contrast with *physics student* in the antecedent since it does not preclude any part of the denotation of the antecedent constituent.

Wagner proposes that the condition on an antecedent for a focus construction should be strengthened so that instead of merely entailing the $\exists$ F-Closure of the focus configuration, the antecedent must be a *local exclusive alternative* of the focus construction. This is formalized using *Universal Closure*, analogous to Schwarzschild’s Existential Closure, in cases where the constituent does not denote a proposition, and using an exhaustivity operator in cases where it does.

Informally, the notion of local exclusive alternative builds on Schwarzschild’s model by mandating that in order to shift prominence away from the default position, the focus configuration under consideration and its antecedent must be mutually exclusive. STUDENT in (125) does not qualify to host prominence because its denotation does not preclude *physics student* in the antecedent; in other words, *physics student* is not an exclusive alternative for STUDENT. Likewise, *someone shot Smith* cannot be its own alternative, since not only is it not excluded by *someone shot Smith*, it is also entailed by it.

Wagner argues that in addition to accounting for data like (124) and (125), his account is

more parsimonious than previous theories of emphasis licensing. This is because theories like Schwarzschild must posit an additional rule to account for cases like (124). In this example, all of the constituents are given, because an identical clause appears as an antecedent. This means that any element of the clause should be able to be emphasized, as any corresponding $\exists F$ -Closure would be entailed by the antecedent. However, when this configuration arises, only the final position is actually observed to host emphasis. Schwarzschild proposes that “default” prominence must hold in such cases, but no theoretical explanation is given for why this particular emphasis configuration arises under these conditions. Wagner, in contrast, also makes use of the notion of “default” emphasis, but by defining a set of conditions under which emphasis can be shifted away from the default position, and in particular one that cannot apply when a constituent is all-given, he avoids the need for an additional rule accounting for the default emphasis pattern in these cases.

Turning to the question of nonidentity in emphasis licensing, Wagner makes liberal reference to the mechanism of accommodation in laying out his theory. He puts forward examples like (126), where a constituent is deaccented despite having no antecedent correlate.

(126) Father: What’s wrong?

Mother: I thought good fiction was valued in this country. Well, I just found out that a SUPERMAN comic *made the bestseller list*.

(Wagner, 2012)

Wagner proposes that this should be understood as accommodation of an antecedent with similar form, e.g., *A serious novel made the bestseller list*. Thus, Wagner’s account parallels those of Tancredi and Fox in requiring accommodation of a covert antecedent to license deaccenting of (at least some) antecedent-nonidentical material.

5.2.3 *Summary*

In summary, this section has outlined two theoretical approaches to accent licensing and anaphoric deaccenting. Each of these is designed to account for some or all of the following types of deaccentable constituents: i) antecedent-identical material, ii) material that co-refers with an antecedent, iii) material that is entailed by an antecedent, modulo existential closure (minimally including synonyms and hypernyms), iv) material that is reasonably concluded from the antecedent based on broader knowledge of the context or the world, and v) material that can trivially be realized without accent in any discourse.

The first approach posits a grammar that directly licenses deaccenting on material that is given or inferable. This includes accounts that give structural definitions of given material with respect to previous linguistic material (e.g., Schwarzschild, 1999) as well as work that treats givenness marking more holistically as a strategic discourse move (e.g., Ladd, 1980; Baumann and Riester, 2012). Many of the proposals assume that the licensing of anaphoric deaccenting should be integrated into a larger focus-theoretic account of emphasis licensing (Ladd, 1980; Rooth, 1992a; van Deemter, 1994, 1999; Selkirk, 1995; Schwarzschild, 1999), but others argue that givenness marking should be kept theoretically separate from focus marking, although the two systems must interact (Büring, 2016). There are still other approaches which argue that givenness marking should replace focus marking altogether. (Sauerland, 2005).

The second approach imposes a grammatical requirement for more or less exact identity between deaccented constituents and their linguistic antecedents. These approaches allow for deaccenting under nonidentity via an extragrammatical mechanism. Listeners are argued to assume rational communication on the part of their interlocutors and accommodate a compatible alternative antecedent when they encounter ungrammatical deaccenting. This proposal is spelled out in detail by Tancredi (1992) and Fox (2000) and such a mechanism is also assumed by Wagner (2012).

These different approaches instantiate at least two fundamentally different theoretical

statuses for the deaccenting of nonidentical material. On the one hand, the grammatical account tends to treat such deaccenting as grammatical and on par with the deaccenting of identical material. On this view, canonical deaccenting under identity is a subset of the set of all grammatical deaccenting, as repeated material is trivially given or entailed. On the other hand, the accommodation account posits that deaccenting nonidentical material is ungrammatical, and requires an extragrammatical operation such as accommodation to be marked as acceptable.

The investigation of deaccenting in the chapters that follow provides not only a novel empirical characterization of the prosodic status of antecedent-nonidentical material, but also a new test for the theory of accent assignment. The grammatical account of deaccenting under nonidentity predicts a more or less uniform empirical status for deaccented identical and nonidentical material (modulo their need to define what nonidentical material counts as “given;” variation in the definition of given material can lead to different predictions about what material should be deaccentable). By contrast, the accommodation account leaves room for “gaps” in the empirical behavior of the two types of material. The systematic experimental investigation discussed in the next two chapters reveals substantial differences between the two in both production and perception, suggesting that the accommodation account may provide a better characterization of the interaction of grammar and extragrammatical processes in licensing deaccenting.

To close, it is worth noting some oversimplifications in both the treatment of the literature and the empirical landscape that will be made for the sake of limiting the dissertation’s scope. First, the discussion above groups prior analyses of deaccenting as “grammatical” if they take a focus or givenness approach and as “accommodation” otherwise. However, many of the theories reviewed actually provide a more nuanced view than this. As an example, Schwarzschild (1999) presents a theory of deaccenting that defines GIVENNESS in grammatical (linguistic) terms, but then notes that a speaker could present information as given that is not formally GIVEN and the result would be acceptable as long as the

hearer could accommodate the “insinuated” antecedent. Similarly, Rooth (1992a) leaves the question of whether to account for world knowledge-based inferences in the grammar or as a broader discourse process open for future research. Thus, the discussion of grammatical and accommodation approaches that will follow is intended to address more abstract strategies for addressing deaccenting inside or outside the grammar, and not as a firm conclusion that any individual analysis described above as primarily grammatical or primarily accommodation-driven should be accepted or rejected wholesale.

Second, and relatedly, the production and perception investigations in the next two chapters will largely treat “accessible” or “inferable” material as a monolithic discourse status that contrasts with new and given (repeated) material. However, the examples cited in this chapter have already made clear that this is overly simplistic. For example, there is a strong tendency in the literature to treat co-referential expressions as straightforwardly deaccentable, and several authors have suggested that nouns or referring expressions have a special prosodic status that requires special theoretical machinery to address (Gunter, 1966; Schmerling, 1976; Ladd, 1980; Neeleman and Reinhart, 1998).

As a result, it will be worth considering in future work whether different types of nonidentical deaccenting lie on different sides of a grammatical/accommodation divide. A particularly attractive proposal would hold that co-referential material is grammatically deaccented, while all other nonidentical material is deaccented through an extragrammatical process. Another possibility adjusts the dividing line to propose that co-referential and formally entailed material is deaccented via a grammatical mechanism, while world knowledge-driven and other contextually dependent cases of deaccenting are pragmatic in nature.

However, it is worth noting here that between the current dissertation and Chodroff and Cole’s (2019) study, a rather diverse set of nonidentical material has already been studied with similar results. Chodroff and Cole studied antecedent-nonidentical nouns with the possibility of co-reference, although they only report production results. The next two chapters study both formally entailed verbs and verbs whose meanings are made available

through more informal, world knowledge-driven relationships to an antecedent, but ones that are still primarily lexical in nature. This leaves room for future study of co-reference in perception, as well as more distantly world knowledge-driven inferences (*call a Republican-insult*) as well as investigation of antecedentless deaccenting, but the similar results between Chodroff and Cole's study and the current dissertation tend to suggest a unified analysis for all deaccented material. Subsequent discussion of the deaccenting studies from the next two chapters will return to these issues.

CHAPTER 6

PRODUCTION OF DEACCENTING UNDER IDENTITY AND INFERENCE

The previous chapter reviewed several theoretical accounts of the licensing of accentuation and deaccentuation. Among other considerations, each model aims to account for two observations pertaining to deviance from default sentential prosody due to deaccenting (stress shift). First, a constituent that would be expected to carry a pitch accent under default prosody can be deaccented when it is identical to a constituent that appears in a recent linguistic antecedent, like the second *saw* in (127).

(127) Bill saw Alice, and then JOHN *saw* SUE.

Second, certain constituents that are “given” in some way with respect to a component of the linguistic antecedent are also claimed to be deaccentable. Impressionistically, several different relations between antecedents and targets are argued to license deaccenting. The target and the antecedent can co-refer, as in (128). The antecedent can entail the target (modulo existential closure), as in (129). The antecedent and target can be linked by more informal relations, as in (130), where *insulted* can be deaccented when the discourse participants take *x insulted y* to be a reasonable conclusion from *x called y a Republican*. Finally, material can be deaccented if the meaning is made salient in the nonlinguistic context, as in (131).

(128) A: Did you see Dr. Cremer to get your root canal?

B: Don’t remind me. I’d like to *strangle the butcher*.

(Büring, 2007)

(129) Bach wrote many pieces for viola. He must have LOVED *string instruments*.

(van Deemter, 1999)

(130) First, John called Mary a Republican. Then, BILL *insulted Mary*.

(Lakoff, 1968)

(131) [Hearer cocks their head to one side as if listening for a faint or distant noise.]

I *heard it*, TOO.

(Rochemont, 1986)

The theoretical accounts can be divided into two groups based on their approach toward unifying these two observations. The first group of models, referred to as the “grammatical account,” focuses on identifying a single rule characterizing constituents that can felicitously be deaccented. This rule is formulated to account for the felicitous cases of deaccenting under inference, while the cases of deaccenting under identity are automatically subsumed as a subset of the cases generated by the rule (for instance, because the constituent’s existential closure entails itself). These models include Rooth’s (1992a) proposal to include an entailment relation in the semantics of his $\sim$ operator, Schwarzschild’s (1999) requirement that the $\exists$ F-Closure of a deaccented constituent be entailed by that of an antecedent constituent, and those models that rely on notional givenness to license deaccenting, such as Rochemont’s (1986) c-construability and approaches that uniformly mark repeated and inferable material as “given” (van Deemter, 1994, 1999; Sauerland, 2005; Baumann and Riester, 2012; Büring, 2016).

The second class of account, termed the “accommodation account,” provides separate licensing mechanisms for the deaccenting of identical constituents versus that of inferable constituents. Generally, this approach holds that deaccenting is canonically (i.e., grammatically) licensed by identity between a linguistic antecedent and the target constituent. Deaccenting licensed by inference, then, relies on the accommodation of a covert antecedent, reasonably concluded from the present linguistic and nonlinguistic context as well as the discourse participants’ world knowledge, that contains a correlate identical to the target deaccented constituent. Accounts taking this tack include Tancredi (1992) (“conversational implicature”), Fox (2000), Wagner (2012), Schwarzschild’s (1999) observation that accommo-

dation is possible when the speaker chooses to present non-GIVEN information as given, and possibly Rooth's (1992a) suggestion that deaccenting under nonidentity should be handled at the "discourse level."

While these two classes of model are intended to account for the same impressionistic data, it is not clear that they make the same predictions for the felicitousness of deaccented identical and inferable constituents. Consider first the grammatical account. In the interest of theoretical parsimony, these models use the same machinery to derive the acceptability of deaccented inferable material as they do for deaccented identical material. What's more, these models often include overt constraints designed to mandate deaccenting whenever it is licensed, such as Schwarzschild's AVOIDF and Büring's MAXIMIZE BACKGROUND. Because these models give a uniform account of deaccenting, they also imply uniform felicitousness of both types of deaccenting. That is, with an antecedent of *hugged*, deaccenting of a subsequent *hugged* is mandatory, but so too is deaccenting of a subsequent *embraced*, since it is inferable from *hugged*. Provided that speakers and listeners are aware of the inferencing relation, there is no clear theoretical motivation for a difference in felicitousness between deaccented *hugged* and deaccented *embraced*.

By contrast, since the accommodation class of models makes reference to two different licensing mechanisms for deaccenting, it leaves open the possibility for felicitousness differences between deaccented identical and inferable constituents. These models cast deaccenting of antecedent-identical constituents as the canonical case; presumably, deaccenting of identical constituents remains mandatory under these models. Deaccenting of inferable constituents, however, requires another operation, schematized as accommodation of a covert antecedent. Proponents of these models clearly did not mean to suggest that this accommodation operation is highly costly, since they rely on it to account for sentences that they mark categorically as acceptable. Nevertheless, the accommodation approach opens the possibility that the licensing of deaccenting on antecedent-identical versus inferable constituents works differently in practice – for instance, deaccented inferable constituents might take longer for speakers

or listeners to process, or they might be rated as slightly less felicitous than deaccented identical constituents.

A potential confound for the prior literature on deaccenting licensing is the fact that deaccenting of antecedent-nonidentical has received minimal systematic attention. While the canonical mapping between repetition and deaccenting is well established, the relationship between a constituent's inferability and deaccenting is based on a small number of introspective judgments. As such, a major goal of this dissertation is to conduct perception studies collecting native speakers' judgments of the felicitousness of deaccented constituents in environments that make them discourse-new, discourse-old, or inferable. A rigorous investigation of the relative naturalness of deaccenting under identity versus inference may help shed light on whether the grammatical or accommodation approach to deaccenting licensing provides a better account of these phenomena. These perception studies are the subject of the next chapter. However, these perception studies require audio-recorded stimuli. While it would be possible to construct these stimuli directly for use in perception studies, this requirement also points to an opportunity to explore the production of deaccenting, with subsequent use of recorded production responses as stimuli in a perception study.

Thus, the topic of the present chapter is a production study that explores the production of prosodic emphasis on constituents of varying discourse statuses. The main goal of this study was to collect naturalistically produced recordings that could be manipulated for later use as perception study stimuli. However, the production task also manipulates the discourse status of certain constituents the participants produced in order to explore how likely speakers are to deaccent antecedent-identical versus inferable constituents in production. The paradigm presented here and in the next chapter focuses on the production of verbs that are new, inferable, or repeated relative to a linguistic antecedent. Verbs and verb phrases have received less attention in the deaccenting literature than noun phrases, which are suggested to participate in additional deaccenting-licensing operations such as co-reference. However, verb phrases are argued to be deaccentable under inference, as in (130). In addition, a sys-

tematic study of verbs here nicely accompanies Chodroff and Cole’s (2019) production study of nouns, one of the only other systematic studies of deaccenting under inference.

It is worth mentioning one major caveat before discussing this production study. Participants in this study read sentences from a script and did not produce any sentences that they planned on their own. The “inferable” constituents they read are made highly available by the linguistic antecedent, as will be shown in Experiment 8. Nevertheless, it is possible that participants in the production study were unaware of the critical inferencing relations - for instance, because they were performing a relatively shallow task of reading the script aloud without fully accessing its meaning. Because of this possibility, any lack of deaccenting observed on inferable constituents should not be taken as evidence that such deaccenting is nonexistent or rare in real-world, naturalistic speech. The perception studies presented in the next chapter will be treated as more conclusively describing the felicitousness of the various forms of deaccenting; the ecological validity of the various experiments will be discussed more at the end of this chapter and at the beginning of the next chapter.

Three studies are presented in this chapter. First, both the production and perception studies rely on the use of antecedent-target pairs of verbs that are string-nonidentical, but where the meaning of the target verb is made highly available from the meaning of the antecedent verb (referred to as “inferable,” although the target is not strictly entailed in all items). Since notions of both formal and informal entailment can be difficult to pin down for verbs, a norming study, Experiment 8, was carried out to elicit native speakers’ intuitions of the strength of the inference relations linking potential stimuli. The results of this study were considered in the selection of stimuli for the subsequent studies.

Experiment 9 is the main production study. In this experiment, participants read aloud sentences of the form *SVO and SVO*. Critically, the second verb could be discourse-new (unrelated or contrastive compared to the first verb), repeated (identical to the first verb), or inferable from the first verb (either via formal entailment or general world knowledge akin to Rooth’s (1992a) “implicational bridging”). The analysis for Experiment 9 focuses on

comparing the phonetic correlates of emphasis collected from new, inferable, and repeated verbs to determine whether inferable verbs are generally pronounced similarly to discourse-new verbs, similarly to repeated verbs, or with a unique prosodic realization.

Finally, Experiment 10 explores the possibility that the phonetic analysis from Experiment 9 failed to capture the impressionistic (phonological) emphasis status of the various constituents recorded in the production study. In this experiment, participants listened to clipped recordings of the Experiment 9 responses such that they were not aware of the constituents' discourse status, and provided a forced-choice rating of the critical constituents as "emphasized" or "not emphasized."

To briefly preview the results, Experiments 9 and 10 largely failed to detect evidence that inferable verbs were deaccented, whereas repeated (antecedent-identical) verbs were reliably deaccented. This constitutes an empirical challenge for the literature on deaccenting, which largely assumes that deaccenting of inferable material is licensed or even mandatory. The results are argued to be more compatible with an accommodation approach to emphasis licensing, as the grammatical account has difficulty accounting for the fact that identical and inferable constituents have very different likelihoods of being deaccented. However, due to the caveat mentioned above that the participants did not plan their own utterances, this evidence is not taken as conclusive, and the exploration of the felicitousness of the two types of deaccenting continues in the subsequent chapter.

6.1 Experiment 8: Norming verb inferability

The primary task in the production study (Experiment 9) was for participants to produce sentences containing verbs with a variety of different discourse statuses – discourse-new, repeated, and critically, inferable. However, during stimulus construction, it was noted that it can be difficult to reach a consensus on what pairs of antecedent and target verbs constitute cases where the second verb is "inferable," particularly by entailment. For instance, while acts of *hugging* may be a subset of acts of *embracing*, there may be subtle semantic or

pragmatic distinctions in the meanings of the words that mean that *x hugged y* does not actually entail *x embraced y*.

In order to account for this variability, Experiment 8 was carried out to identify pairs of verbs with sufficiently strong inferencing relations. Essentially, the goal of the experiment is to find target verbs that are highly inferable from a particular non-identical antecedent verb, even though all speakers may not agree about the type of inference relating them (e.g., formal versus “informal” entailment).

The experiment elicited participants’ ratings of how likely they were to think that certain inferable and discourse-new verb actions were to be the case given a particular antecedent. The results led to the selection of a set of stimuli with highly inferable target verbs in the “inference” condition and high contrast between the inferability of the inference verbs and of the discourse-new verbs.

6.1.1 *Design and materials*

Two sets of stimuli were tested. For the first, the intended inferencing relation linking the antecedent and target verb in the “inference” condition was formal entailment – that is, it was intended that the existential closure of the antecedent verb entailed that of the target verb. An example of such a pairing is *hugged* and *embraced*. If hugging acts are a subset of embracing acts, then *x hugged y* entails *x embraced y*. According to the deaccenting literature, this would mean that an antecedent of the form *x hugged y* can license a follow-up clause with a deaccented *embraced* – because the existential closure of *embraced* is entailed, the meaning of *embraced* is available in the context, or a covert antecedent containing *embraced* can be accommodated, depending on the model.

In the second set of stimuli, the inferencing relation was more informal, with the second verb denoting an action that might generally be considered reasonable given that the denotation of the first verb is the case. In these items, the second verb is not strictly “inferable” from the first. Rather, the goal is that if a listener were to encounter a deaccented instance of

the target verb with a particular antecedent, they might reasonably conclude that the presence of the second verb in the discourse context was unremarkable based on the presence of the first verb. This set of stimuli is inspired by what Rooth (1992a) calls the “implicational bridging” cases, e.g., where *insult* is treated as unremarkable due to the presence of *call a Republican* in the context.

An example of a verb pair with this “implicational bridging” type of relationship is *charmed* and *seduced*. *x charmed y* certainly does not formally entail that *x seduced y*. Nevertheless, the two verbs are highly semantically related, and it is likely to be unsurprising that *seduced* holds given that *charmed* is the case. Thus, *charmed* might serve as an antecedent for a deaccented *seduced*, as long as a listener is willing to count the relation between the verbs as one of informal entailment, take *charmed* as a semantic antecedent for *seduced*, or accommodate a covert antecedent containing *seduced*, depending on the theoretical model.

For both the entailment relation and the implicational bridging relation, 30 triples of past-tense transitive verbs were constructed. Within each triple, one verb was the antecedent, and two verbs were potential targets. One target was *discourse-new* relative to the antecedent, meaning it was intended not to be inferable from the antecedent verb (but see Appendix I for discussion of this condition, which may consist of a mix of contrastive and new verbs). The other target was *inferable*, meaning it was intended to be entailed by the antecedent (modulo existential closure) or made available under implicational bridging.

The goal of the experiment was to elicit participants’ ratings of the likelihood that the existential closure of the target verb could be inferred from that of the antecedent verb. For each trial, the antecedent verb and one target verb were paired with the same proper name subject and object to form two sentences of the form *subject verb object (SVO)*. The sentences were embedded in the question, “Suppose you know that [antecedent sentence]. How likely do you think it is that [target sentence]?” The question was accompanied by a 7-point Likert scale, where 1 was labeled as “least likely” and 7 was labeled as “most likely.” Different proper names were used for each item to avoid participants remembering

any actions that might have previously been attributed to a certain name on a previous trial.

Table 6.1 shows sample antecedent and target stimuli for an entailment item and an implicational bridging item. The full set of verb triples is shown in Appendix I.

Entailment	Antecedent	Brandon hugged Katherine.
	Target, discourse-new	Brandon rebuffed Katherine.
	Target, inferable	Brandon embraced Katherine.
Imp. bridging	Antecedent	Steven charmed Andrew.
	Target, discourse-new	Steven offended Andrew.
	Target, inferable	Steven seduced Andrew.

Table 6.1: Sample stimuli for Experiment 8

Although the studies below contain a condition where the verbs in the antecedent and target clause were identical, this condition was left out of the norming study due to concerns that participants would find the task of rating the likelihood of a target sentence given a string-identical antecedent discursively odd. It is assumed that the mean likelihood ratings for such trials would be very close to 7, modulo experimental noise.

6.1.2 Procedure

The experiment was conducted using the Amazon Mechanical Turk platform. Participants were recruited using a HIT description noting that they were required to be over the age of 18 and a native speaker of English to participate in the experiment. Participants were also required to have a United States IP address to participate. After accepting the HIT, they were linked to the experiment, which was hosted on Ibex Farm, where they provided informed consent, completed a demographic survey, completed the experiment, and then received a survey code. After completing the Ibex Farm portion of the experiment, they returned to the AMT site and entered the survey code to indicate experiment completion and receive payment. Participants received 1 US dollar for participating in the experiment.

After providing informed consent and completing the demographic survey, participants read the instructions for the experiment. The instructions familiarized the participants with

Use the scale to answer the question, where 7 represents *very likely* and 1 represents *very unlikely*.

Suppose you know that

Raymond despised Gregory.

How likely do you think it is that

Raymond hated Gregory?

(Unlikely) (Likely)

Use number keys or click boxes to answer.

Figure 6.1: Sample screen from Experiment 8

the task and warned them that if they were unsure of their answer on a particular trial, they should answer with their first instinct rather than overthinking their response. Following this, participants completed two practice trials to verify that they understood the setup and the method for entering a response (clicking the number rating on the screen or pressing the corresponding number key on their keyboard).

Two lists were created for the experiment. In one list, half of the entailment items and half of the implicational bridging items featured the discourse-new verb in the second sentence, while the remaining items had the inferable verb in the second sentence. In the other list, the discourse status of the second verb was reversed for each item. Participants were assigned to the lists in sequential order of starting the experiment, so that the first participant to begin was assigned to the first list, the second participant to the second list, and so on. The order of presentation of the critical items was randomized for each participant.

A screenshot showing the layout of one trial of the experiment is shown in Figure 6.1.

6.1.3 *Participants*

60 participants took part in the experiment. Of these, 24 were female and 36 were male. The mean age of the participants was 33.2 years. The data from 5 subjects was excluded from analysis because their mean reaction time was under 2000ms, suggesting that they were quickly pressing response keys to complete the experiment without fully attending to the task.

Among the participants whose data was analyzed, the mean time to complete the AMT portion of the experiment was 10 minutes, 16 seconds. Participants were compensated with USD 1.00 for completing the experiment.

6.1.4 *Results and stimulus selection*

On a scale from 1 to 7, where 7 is the most inferable, the mean ratings for the new and inferable verbs in the two sets of items are as follows: entailment/inferable: 6.49; entailment/discourse-new: 2.47; bridging/inferable: 5.84; bridging/discourse-new: 2.43. The mean results for each item (verb triple) are shown in Appendix I.

Impressionistically, the high means of the inferable items indicates that the stimulus construction was largely successful, i.e., that the existential closures of the target inferable verbs are highly inferable from the existential closures of the corresponding antecedent verbs. Further, there is a sizeable gap in inferability between the inference target verbs and the discourse-new verbs. This is important because the discourse-new condition serves as a baseline for canonical/felicitous accenting, and the experiments below aim to detect deaccenting of inferable constituents by directly comparing the behavior of inferable verbs to that of discourse-new verbs. Thus, a larger gap in inferability between these two conditions increases the likelihood of the experiments detecting deaccenting-under-inference behavior among the inferable constituents.

A final set of 24 entailment items and 24 implicational bridging items was selected by discarding the 6 items in each set that had the highest inferability score for the discourse-new

condition. This was done to widen the gap in inferability between the inferable verbs and the new verbs, as these two conditions are directly compared in the subsequent experiments to determine whether inferable verbs are treated differently from discourse new (i.e., can be deaccented). The 48 verb triples that were selected were used in Experiments 11 and 12 and are indicated in Appendix I. For the final set of 48 items (i.e., with the 12 low-inferability scoring items discarded), the mean inferability scores are as follows: entailment/inferable: 6.49; entailment/discourse-new: 2.12; bridging/inferable: 5.78; bridging/discourse-new: 2.15.

Experiment 9, the production study, featured careful matching of the syllable and lexical stress patterns of the target-clause stimuli. For this study, a subset of 6 entailment items and 6 bridging items was selected from the final set of 48 experimental items. The most common stress pattern among the target verbs was iambic stress. From the implicational bridging stimuli, the 6 items featuring an iambic target verb were selected. Among the entailment items, there were only 5 items with iambic target verbs, so these items plus a sixth item with the target verb *admired*, which was impressionistically judged to be the next closest verb to an iambic pattern, were selected. The 12 verb triples that were selected for use in Experiment 9 are indicated in Appendix I. For the 12 items used in Experiment 9, the mean inferability scores are as follows: entailment/inferable: 6.56; entailment/discourse-new: 1.86; bridging/inferable: 5.49; bridging/discourse-new: 2.19.

6.2 Experiment 9: Production of deaccenting

With Experiment 8 having confirmed that the constructed verb triples exhibit the intended inferability relations, the main production study of the chapter could be conducted. Experiment 9 is a production study designed to elicit and record naturalistic tokens of constituents under various discourse statuses. The main manipulation is of the status of a particular verb, which can be discourse-new, repeated, or inferable (via entailment or implicational bridging) relative to the content of the previous clause.

The main goal is to analyze several phonetic correlates of emphasis (mean intensity,

mean f_0 , duration) on the nucleus of the manipulated verb constituent. Canonically, in a sentence of the form *SVO and SVO*, a discourse-new second-clause verb should carry a pitch accent, while a repeated second-clause verb should be deaccented. A comparison of the phonetic characteristics of the inferable verbs to these baselines will indicate whether the inferable verbs are accented, deaccented, or have their own unique phonetic realization. The results will inform the theoretical accounts of the licensing of prosodic emphasis, which make (potentially) different predictions about whether repeated and inferable constituents can exhibit different prosodic behavior in production.

6.2.1 *Design and materials*

Participants were recorded reading three-sentence paragraphs aloud. The second of the three sentences was the critical sentence of interest, whereas the first and third sentences were used as a carrier to reduce the impact of any edge effects on the prosodic realization of the second sentence (Pierrehumbert and Hirschberg, 1990; Silverman et al., 1992).

The critical second sentence was invariably of the form *SVO and SVO*, where *S* represents a subject, *V* represents a transitive verb, and *O* represents an object. The subjects and objects were proper given names. The main question of the experiment is how the prosodic realization of the second SVO clause varies in response to changes in its relationship to the first SVO clause, and in particular how the emphasis status of the second-clause verb varies depending on its relationship to the first-clause verb.

The experiment featured a 2 (object status) x 3 (verb status) design manipulating the discourse status of the constituents in the second clause. In the stimuli, both manipulations were manifested in the first clause of the *SVO and SVO* sentence. The second clause was held constant so the phonetic values extracted from it could be directly compared to one another, but the underlying discourse status of each constituent actually varied according to the constituents that appeared in the first clause.

The second-clause object could be *new*, meaning it was not the same name as the first

clause object, e.g., *Eliza delayed Shirley, and Will amused Mary*. Alternately, the second object could be *old*, meaning the objects in the two clauses were identical, e.g., *Samantha delayed Mary, and Will amused Mary*.

The goal of this object manipulation was twofold. First, it provides an opportunity independent of the critical verb manipulation to check that the experimental and analytical paradigm is sensitive to differences in the realization of prosodic emphasis. The expectations for the prosodic realization of the second-clause object are fairly straightforward: a fully-discourse new object should carry a pitch accent, while a repeated object should be deaccented. Thus, if the analysis used for this experiment successfully detects such an effect, it suggests that the same techniques should be valid for the analysis of the critical verb data.

Second, manipulating the discourse status of the object increases the number of sentential configurations available for analysis. When the object is new and canonically accented, it should carry a nuclear pitch accent, as it is the most embedded constituent in the clause. This means that the verb can carry at most a prenuclear pitch accent. By contrast, when the object is old and canonically deaccented, the verb is the most embedded remaining constituent and is in a position to possibly carry a nuclear pitch accent. This allows for analysis of two different contrasts on the verb: between an unaccented verb and a non-nuclear accented verb when the object is new, and between an unaccented verb and a nuclear accented verb when the object is old. Note that there is little expectation that verbs in prenuclear position will exhibit strong prosodic differences, as prior research has suggested an unreliable mapping between information structure and accent in prenuclear position (Chodroff and Cole, 2018).

In addition to the object manipulation, there was a three-way manipulation of the discourse status of the verb in the second clause. The verb could be *repeated*, meaning it was the same verb as in the first clause, e.g., *Tiffany amused Shirley, and Will amused Mary*. It could also be *new*, meaning that the second verb was neither identical to the first nor inferable from it, e.g., *Eliza delayed Shirley, and Will amused Mary* (but see the note in

Appendix I on whether the “new” verbs were new or contrastive). Finally, the second verb could stand in an *inference* relation with the first, e.g., *Patty entertained Shirley, and Will amused Mary*.

As in Experiment 8, there were two groups of items differentiated by the nature of the *inference* relation. In one half of the items (1-6), the inference relation was entailment, meaning the $\exists$ -closure of the first verb entails that of the second (as normed in Experiment 8). In the other set of items, the relation was one of implicational bridging, meaning that the $\exists$ -closure of the first verb informally supports the possibility that the $\exists$ -closure of the second verb is the case.

Recall that the items included in this experiment are those items where the target inference verb was an iamb (plus the item with the target *admired* among the entailment verbs) that were not discarded as having an inferability score for the discourse-new verb that was in the top fifth of the items in the relevant inference relation group. The mean inferability scores for the included items as measured in Experiment 8 (on a 7-point scale where 7 is the most inferable) are: entailment/inferable: 6.56; entailment/discourse-new: 1.86; bridging/inferable: 5.49; bridging/discourse-new: 2.19.

The goal of the verb manipulation is to determine whether verbs that are inferable from the first clause are realized with or without a pitch accent. The new and repeated conditions act as a baseline for comparison, as new verbs canonically are expected to be accented and repeated verbs are expected to be deaccented.

The subject of the second clause was always discourse-new. This is because the discourse status of the second subject was not expected to affect the prosodic realization of the critical verb. Since the subject is less embedded than the verb, whether the subject is accented does not affect the possibility of the verb carrying a nuclear versus prenuclear pitch accent. In addition, adding a new-old manipulation for the subject to the current design would result in a cell where the second subject, verb, and object were all identical to the first clause (e.g., *Will amused Mary, and Will amused Mary*), which is pragmatically odd compared to the

other sentences in the experiment and may have affected the performance of the participants.

In order to further control the discourse status of the constituents within the critical second sentence, the first sentence of the carrier paragraph did not use any of the proper names or verbs from the critical sentence (in fact, proper names were not used in any of the context sentences). Care was taken to avoid providing any information in the first sentence that might make any of the proper names or verbs in the second sentence inferable.

In order to reduce the incidence of phonetic artifacts between different productions, the syllable structure of the critical second SVO clause was constant for all items with one exception. The second-clause subject was always a monosyllable. The object was always a trochee, i.e., a two-syllable word consisting of a stressed syllable followed by an unstressed syllable. The verb was always an iamb, i.e., a two-syllable word consisting of an unstressed syllable followed by a stressed syllable. The exception is the verb *admired* in item 5, which is likely realized by most speakers of North American English as a three-syllable word. This verb was included because there were only five iambic verbs in the entailment group remaining after Experiment 8. For *admired*, measurements from the stressed [aɪ] nucleus are the ones reported alongside the stressed-syllable results for the other verbs, while those from the [ɪ]/[ɜ] nucleus were disregarded.

In addition to the syllable constraints described above, the subject and object proper names used in the second clauses invariably consisted only of sonorants and vowels to facilitate the extraction of phonetic information such as f_0 from the recordings.

To further reduce the variability between items, the number of syllables before the onset of the critical clause was also held constant. The first carrier sentence always consisted of exactly 12 syllables. The first SVO clause always consisted of 7 syllables. The object of this clause was always two syllables; it was either identical to the second-clause object (old object), or a different two-syllable name (new object). The first verb must vary to condition the discourse status of the second verb, and as such, it was not practical to hold the syllable count of the first verb constant. Instead, the first-clause subject was allowed to vary (and in

First sentence: My brother’s birthday party got rowdy last week.

<i>Verb status</i>	<i>Object status</i>	<i>Second sentence</i>
New	New	Sebastian punished Justin, and Nan surprised Amy.
New	Old	Gabriel punished Amy, and Nan surprised Amy.
Inference	New	Bradley astounded Justin, and Nan surprised Amy.
Inference	Old	Ethan astounded Amy, and Nan surprised Amy.
Repeated	New	Julian surprised Justin, and Nan surprised Amy.
Repeated	Old	Benjamin surprised Amy, and Nan surprised Amy.

Third sentence: At least there’s leftover cake to enjoy.

Table 6.2: Sample Experiment 9 stimuli with entailment as the inferencing relation.

First sentence: Things were chaotic in my house the other day.

<i>Verb status</i>	<i>Object status</i>	<i>Second sentence</i>
New	New	Isabella dressed Felix, and Ray harassed Larry.
New	Old	Alexander dressed Larry, and Ray harassed Larry.
Inference	New	Mohammed annoyed Felix, and Ray harassed Larry.
Inference	Old	Elliot annoyed Larry, and Ray harassed Larry.
Repeated	New	Avery harassed Felix, and Ray harassed Larry.
Repeated	Old	Everett harassed Larry, and Ray harassed Larry.

Third sentence: Luckily, the kids all made it to school on time.

Table 6.3: Sample Experiment 9 stimuli with implicational bridging as the inferencing relation.

fact, was different for each verb/object combination for each item) so that the total number of syllables between the first subject and first verb was always 5. Thus, including the *and* linking the two SVO clauses, the onset subject of the critical second SVO clause was always the 21st syllable of a production trial.

Example carrier and critical sentences for an item with entailment as the inferencing relation are shown in Table 6.2. Example sentences for an item with implicational bridging as the inferencing relation are shown in Table 6.3. The full set of experimental stimuli are shown in Appendix J.

6.2.2 Procedure

Each participant read all twelve items in all six condition combinations, for a total of 72 three-sentence paragraphs produced. The was divided into four blocks of eighteen trials

each. Between each block, participants performed an unrelated task where they listened to a ten second clip of classical music and provided a rating indicating how much they liked it. This task was intended to give participants a break and prevent excessive boredom with the production task, and participants also had the opportunity to take a break before beginning the next block.

After providing informed consent, participants entered a double-walled sound booth and were fitted with a pair of headphones (used to play the classical music clips) and a Shure SM10A head-mounted microphone connected to a Zoom H4n recorder. The entire experimental session was recorded using Audacity at a sampling frequency of 44,100 Hz.

The experiment was presented on a computer screen using PsychoPy (Peirce et al., 2019). Before beginning the task, participants read instructions on how to complete both the reading phase and the music clip phase. For the critical reading trials, participants were told that the entire paragraph would display on the screen. Participants were told to read the paragraph in its entirety and plan how they would read it before beginning to speak. This was meant to encourage participants to plan their production of the second sentence instead of reading the words “as they came,” as they were impressionistically expected to be more likely to produce naturalistic emphasis patterns if they were aware of identity and inference relations from previous exposure to the sentence. In addition, in order to elicit clear speech, participants were told to imagine that they were speaking to a listener who was hard of hearing.

Before beginning the first block of production trials, participants completed one practice production trial and one practice music trial to verify that they were comfortable with the method for advancing through the screen display and providing keyboard responses for the music trials.

Within a production block, a trial consisted of a warning screen telling participants that a paragraph would appear when they pressed the spacebar. Upon pressing the spacebar, the paragraph appeared, at which point the participant read the paragraph to plan their production and then read it again out loud. After finishing the production, participants

pressed spacebar to indicate that they were finished and again saw the warning screen, if another production trial followed, or an indication that they were entering a music phase, if they had just completed the last trial in a block.

In a music trial, participants saw a warning screen telling them that pressing spacebar would begin music playback. When they pressed spacebar, a music clip played, and when the clip ended, participants were asked to rate how much they liked the clip on a scale from 1 to 7 using the number keys on the computer keyboard. Providing this rating returned the participant to a warning screen if another music trial followed or a screen informing them that a reading phase was beginning if a phase boundary followed.

6.2.3 Participants

The participants were 10 native speakers of American English recruited from the University of Chicago campus community or its environs. The majority of the participants were recruited through a departmental subject pool and compensated with course credit. The remainder were recruited through the subject pool or word of mouth and were compensated with USD 10.00.

Five of the participants were female and five were male. The mean age of the participants was 21.9 years. The mean time taken to complete the core part of the experiment (excluding consent, demographic questionnaire, and payment), as measured by the length of the sound file recorded in Audacity, was 34 minutes, 37 seconds.

6.2.4 Data processing

For each participant, the recording of the experimental session made using Audacity was processed using Praat (Boersma and Weenink, 2013). A TextGrid was created for each sound file. The approximate onset and offset of each paragraph-length production trial was manually marked in each TextGrid. Each trial interview was labeled with a transcript of the full paragraph that was produced at that point in the experiment.

Each sound file and TextGrid was forced aligned using the Penn Phonetics Lab Forced Aligner Online Processing System (Yuan and Liberman, 2008), with a supplemental file providing Arpabet transcriptions for names in the transcript that are not in the Penn Forced Aligner dictionary. The forced aligner compares phonetic transcriptions of words in its dictionary to the phonetic properties of the uploaded sound file to automatically segment labeled and transcribed intervals larger than the word level, as found in the accompanying TextGrid, into labeled intervals at both the word and phone level that are aligned with the phonetic signal.

The forced aligner is largely accurate, but there were some stimuli in the experiment that led to incorrect segmentation. For instance, the sequence *Anna and* in the stimulus *Abigail annoyed Anna, and Lynn annoyed Anna* was often mis-segmented due to the acoustic similarity of the end of *Anna* and the onset of *and*. Thus, the sound files and TextGrids were reanalyzed in Praat and the phone boundaries in the critical second SVO clause were manually adjusted to align with the acoustic signal as needed.

Next, the Praat script ProsodyPro (Xu, 2013) was run on the sound files and their corresponding phone-segmented TextGrids. This script automatically extracts values for a variety of prosodic variables for each labeled segment in a TextGrid (i.e., for each phone). The analysis below discusses the results of three of these variables: mean intensity, mean f_0 , and duration. However, ProsodyPro actually gives values for several more variables than this, including some that were noted in the previous chapter as being relevant to the production and perception of emphasis (e.g., peak rather than mean f_0), so it may be worth revisiting some of this additional data in the future.

The outputs of the ProsodyPro script for each participant were combined and analyzed using RStudio. For the analysis presented below, only the results for the vocalic nuclei of the five critical second-clause syllables were analyzed, as vowels are expected to carry the bulk of information regarding a word's prosody. (As noted above, for the three-syllable verb *admired*, the second-syllable readings were taken from the [aɪ] nucleus, while readings from

the third syllable were disregarded.)

For the f_0 results, ProsodyPro gave many spurious readings. This is due in large part to the presence of creaky voice in the recordings, which can interfere with the automatic recognition of f_0 values in Praat. To compensate for this, the sound files were manually reexamined for creaky voice and other spurious f_0 readings using Praat’s “Show Pitch” feature.

For vowels where the f_0 reading yielded by ProsodyPro was judged to be spurious because the vowel was partially modally voiced and partially creaky, or for fully modal vowels with spurious readings, the ProsodyPro values were replaced with values obtained using Praat’s “Get Pitch” feature over the largest interval for which the vowel was fully modal. Such manual corrections were made for 139 out of 3595 total vocalic nuclei, or 3.6% of the data. (Note that there were actually 4000 critical vocalic nuclei, but one critical clause was left out of the ProsodyPro analysis due to experimental error.)

926 nuclei were judged to be fully creaky or otherwise have unrecoverable f_0 measurements, for a total of 25.8% of the data. Of these, 511 were O2 (second syllable of object) readings. This syllable was always lexically unstressed and it always sentence-final, factors which conspire to make creaky voice very likely, so it is unsurprising that f_0 information was rarely recoverable (71.1% of O2 syllables were affected).

The remaining 415 nuclei with unrecoverable f_0 values constitute only 14.4% of the data. Of these, 114 were on unstressed verb syllables (15.9% of V1 readings) and 245 were on stressed object syllables (34.1% of O1 readings). The V1 findings are unsurprising since the syllable is unstressed, while the O1 findings are likely due to the fact that the object is sentence-final and that canonically, 50% of the objects in the experiment are expected to be deaccented.

For the remaining syllables, 54 stressed verb syllables had unrecoverable f_0 information (7.5% of V2 readings), and only 2 subject nuclei did (0.3%).

Those syllables with unrecoverable f_0 values were removed from the f_0 analysis, and the

possible effects of this removal, in proportion to the amount of data that was removed, on the analysis for each syllable should be kept in mind during the discussion below.

To compensate for inter-speaker variation in typical f_0 values, the remaining f_0 readings were z-transformed within subjects so that the data was expressed in terms of its typicality or atypicality for the speaker that produced the particular nucleus. Following this transformation, to account for any additional spurious readings and other outliers, data with z-scores above 2 or below -2 (i.e., the upper and lower 2.5% of data) were discarded.

Speakers were also expected to exhibit variation in their intensity values, so the mean intensity readings were likewise z-transformed within each participant. Again, the upper and lower 2.5% of data were discarded.

For the duration readings, it is less clear that different speakers should be expected to exhibit markedly different values. The analysis below reports the findings for raw duration measures, with those data points falling below percentile 2.5 or above percentile 97.5 for absolute duration discarded.

It is possible that absolute duration values may vary within the experiment due to factors like changing speech rate, relativized duration was also explored during data analysis. The data from one type of relative duration analysis and some remarks on this analysis are shown in Appendix J; the statistical analysis of the relativized duration data did not show any qualitative differences from the analysis of the absolute duration data, so the absolute data is discussed below.

During manual correction of interval boundaries and f_0 readings, 10 trials were found to contain major production errors. The data for all five syllables in these trials were excluded from the analysis of each phonetic correlate (1.4% of trials). For the duration analysis presented below, each of the trials with the same speaker/item combination as one of these error trials was excluded from analysis (7.5% of the data in total). Since the duration data was unnormalized, it was expected that the analysis would be more robust if the design was fully balanced between each condition. That is, for a typical item, each speaker produced

the item in all six verb/object condition combinations. In those cases where there was a production error, the data was removed from all six conditions so that possible word-based effects on duration would remain balanced across the conditions. This was not performed for the intensity or f_0 analyses since those data were normalized within each participant.

6.2.5 *Predictions*

Independent of the critical research question of the experiment, there are a variety of expectations for the results based on established facts about the relationship between discourse status and prosodic emphasis. The manipulation of object status should yield a binary pattern of results on the second-clause object, with discourse-new objects generally carrying a nuclear pitch accent and discourse-old objects being deaccented. This will be an important barometer for determining whether each component of the analysis is suitable for detecting categorical differences in emphasis between different conditions of the experiment.

Turning to the manipulation of verb status, it is canonically expected that discourse-new verbs should carry a pitch accent (i.e., be emphasized), whereas repeated verbs should be deaccented. By extension, then, it is expected that the phonetic correlates of emphasis for new verbs will universally be significantly higher than those for repeated verbs. Thus, new verbs will be assumed to be prototypically emphasized and repeated verbs prototypically de-emphasized, with the two groups serving as the standards of comparison for determining the emphasis status of inference verbs.

It is worth noting that new verbs are expected to carry a nuclear pitch accent only when the following object is discourse-old. When the object of the second clause is discourse-new, any accent on the verb will be pre-nuclear. Thus, it is not clear that the pattern of results on the verbs should be expected to be as clear in the new-object conditions as in the old-object conditions, as prenuclear accent is less predictably realized than nuclear accent in English (Chodroff and Cole, 2018).

Of critical theoretical interest is the emphasis status of inference verbs. The phonetic

correlates for inference verbs will be compared to those for both new verbs and repeated verbs to determine how their prosodic realization compares to that of canonically accented and deaccented verbs, respectively.

Three possible results are anticipated for inference verbs. First, the phonetic correlates of emphasis on these verbs might show no difference from those for new verbs while being significantly higher than those for repeated verbs. This would indicate that both new and inference verbs were accented, while repeated verbs were deaccented. This finding would support theoretical approaches that treat the deaccenting of inferable constituents separately from deaccenting of antecedent-identical material, such as Fox's (2000) proposal that deaccenting of inferable material is due to accommodation of an identical antecedent. By contrast, this finding would be problematic for accounts that provide a unified treatment of identical and inferable material, especially those that suggest that deaccenting of inferable material is highly likely or even mandatory (e.g., Schwarzschild, 1999).

Second, it may be the case that the phonetic correlates for inference verbs show no difference from those for repeated verbs, while being significantly lower than those for new verbs. This would suggest that inference verbs, like repeated verbs, were universally deaccented to the exclusion of new verbs. In contrast to the previous possible finding, this would support accounts providing a unified treatment of identical and inferable material that make deaccenting of inferable material likely or mandatory. This finding would also be compatible in principle with the accommodation account, since it is designed to straightforwardly handle deaccenting of inferable constituents. On the other hand, if accommodation is costly in some way relative to grammatical deaccenting under nonidentity, this account might predict slightly less deaccenting of inferable constituents, in contrast to this pattern of results.

Finally, it may be the case that the phonetic values for inference verbs are both significantly lower than those for new verbs and significantly higher than those for repeated verbs, indicating an intermediate emphasis status for inferable material. Such a finding might be driven by two separate patterns within the data. First, it may be the case that inferable

material was generally pronounced with prosodic realizations in between canonical emphasis and canonical lack of emphasis. This might support a theoretical account that treats identical and inferable material separately, with the licensing mechanism for the deaccenting of inferable material being treated as more marginal for that of identical material or otherwise explaining the incomplete de-emphasis of such material.

Alternately, it might be the case that an intermediate finding for the means of the phonetic correlates for inference verbs is driven by the presence of both canonically accented and deaccented inference verbs in the data – that is, on any given trial, participants chose alternately between fully accenting or fully deaccenting inference verbs. This would require a theoretical account that can explain both patterns of behavior for inference verbs. For instance, inferable constituents might be deaccented when the strength of the inferencing relation passes a certain threshold to make an identical antecedent accommodable, and accented otherwise, or speakers may simply make a random choice between accenting and deaccenting when the constituent in question is inferable.

Relatedly, it is not clear *a priori* whether identical results are expected in the inferable conditions for the entailment items and the implicational bridging items. In the literature, both types of inferencing relations are presented as sufficient to license deaccenting. On the other hand, only the entailment antecedents make the inferable target verbs strictly available, whereas the inferencing relation for the bridging items is more informal and might not be recognized by some participants. This may lead to a finding that entailed inference verbs were deaccented more readily than bridging verbs.

6.2.6 Results

The results for z-transformed intensity, z-transformed f_0 , and duration are shown in Figures 6.2, 6.3, and 6.4, respectively. In these plots, the results are split between the upper and lower frames according to item number. For items 1-6, in the top frames, the inferencing relation in the inference condition is entailment, while for items 7-12, in the bottom

frames, the relation is implicational bridging. The left frames show those conditions where the second-clause object was repeated from the first clause, while the right frames show the conditions where it was new. Within each frame, the left-to-right progression shows the nuclei of the second-clause syllables in order: subject (S); first, unstressed verb syllable (V1); second, stressed verb syllable (V2); first, stressed object syllable (O1); and second, unstressed object syllable (O2).

A visual inspection of the plots suggests two major trends. First, the phonetic values for the stressed object syllable, O1, appear higher when the object is new compared to the corresponding conditions where the object is old. This is expected, since a new object canonically should receive a nuclear pitch accent in an *SVO* clause, with most of the phonetic information conveying emphasis on the stressed syllable, while a repeated object canonically should be deaccented. This apparent trend suggests that the experimental paradigm and data processing methods are sufficiently sensitive to detect differences in prosodic emphasis.

Second, the phonetic values for a repeated verb are generally lower than those for new and inference verbs. This trend appears to be robust on the stressed second verb syllable, V2, when the object is old. The trend is less consistent on the unstressed verb syllable, V1, as well as on V2 with a new object. By contrast, the phonetic values for new and inference verbs do not appear to differ by a substantial margin for any phonetic correlate on either verb syllable. This trend suggests that the experimental paradigm detected emphasis differences on verbs – new verbs were accented and repeated verbs were deaccented, as canonically expected – but that inferable verbs were not deaccented to a degree distinguishing them from new verbs. The results for each phonetic correlate and each syllable are discussed in more detail in the analysis subsection below.

6.2.7 Analysis

The results for Experiment 9 were analyzed separately for each of the five second-clause syllables. Below, the three second-clause constituents are discussed in reverse order: objects,

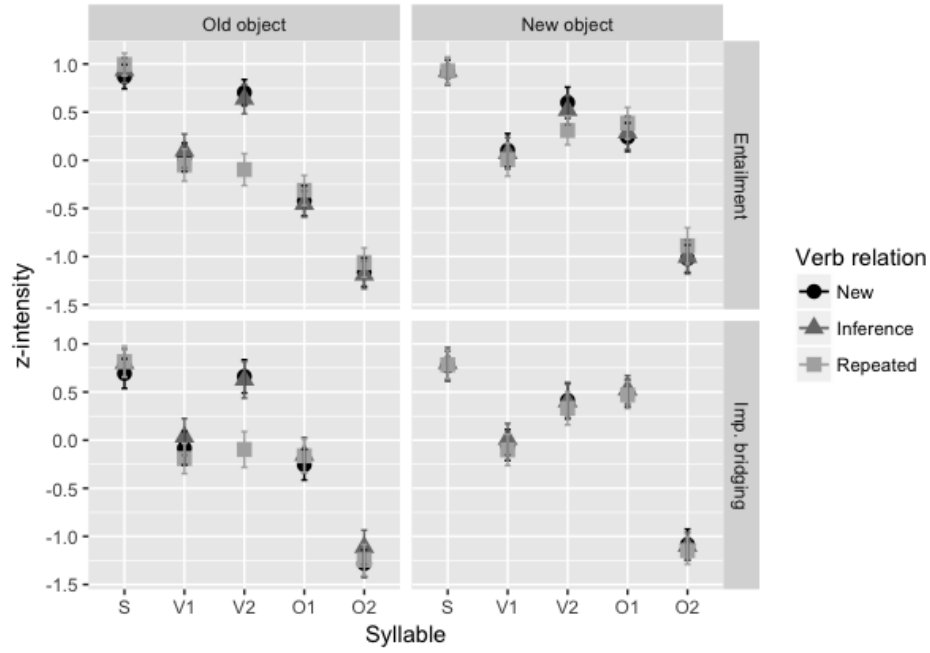


Figure 6.2: Experiment 9 results for z-transformed intensity. Horizontal position: Syllable within critical clause. Horizontal split: Object status. Vertical split: Inferencing relation (items 1-6 versus 7-12). Error bars: 95% confidence interval.

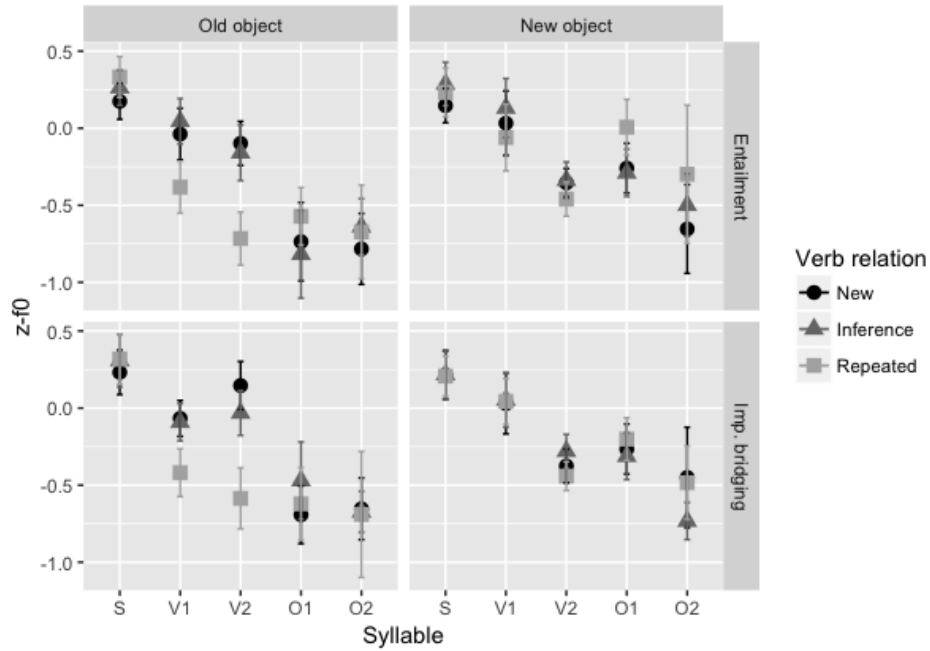


Figure 6.3: Experiment 9 results for z-transformed f_0 . Horizontal position: Syllable within critical clause. Horizontal split: Object status. Vertical split: Inferencing relation (items 1-6 versus 7-12). Error bars: 95% confidence interval.

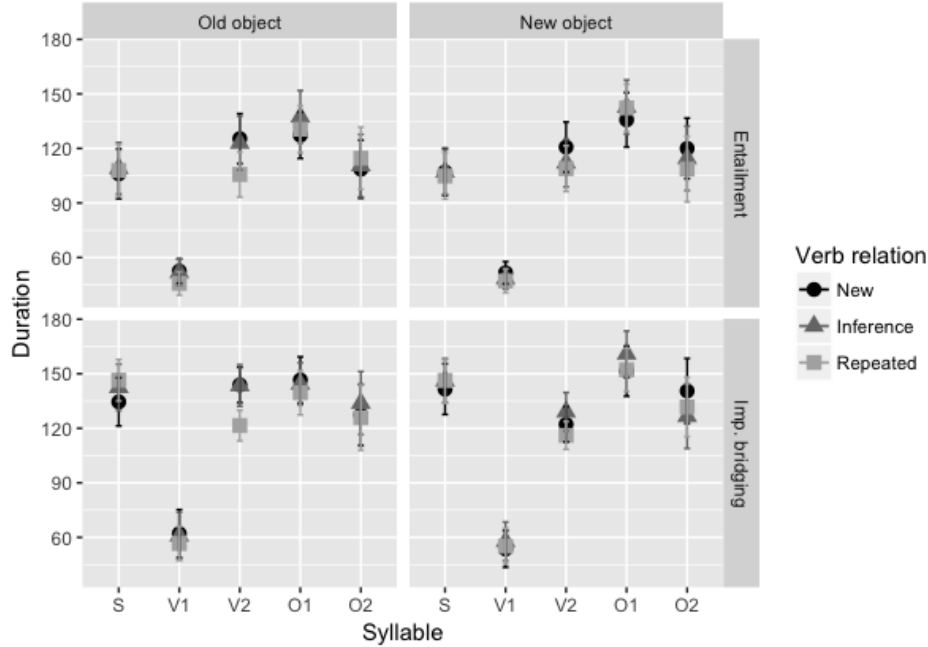


Figure 6.4: Experiment 9 results for duration. Horizontal position: Syllable within critical clause. Horizontal split: Object status. Vertical split: Inferencing relation (items 1-6 versus 7-12). Error bars: 95% confidence interval.

then verbs, then subjects. Objects are discussed first because they serve as a clear, independent test of the experimental paradigm. Canonically, new objects should be accented and old objects should be deaccented, so if the experiment successfully elicited distinguishable prosodic behaviors from the participants, then the analysis should reveal differences in the phonetic correlates between new and old objects.

Verbs are the primary constituent of interest in this experiment, as they were the locus of the critical discourse status manipulation. Subjects are discussed last because there are no predictions of note regarding their emphasis status, as all subjects were discourse-new in this experiment. However, there may be interesting phonetic differences between those subjects presented to bear a nuclear (i.e., followed by deaccented verb and object) versus non-nuclear (i.e., followed by at least one accented constituent) pitch accent.

Objects

The results for z-transformed intensity, z-transformed f_0 , and duration on O1 are shown in Figures 6.5-6.7. For each phonetic correlate/syllable combination, separate linear mixed-

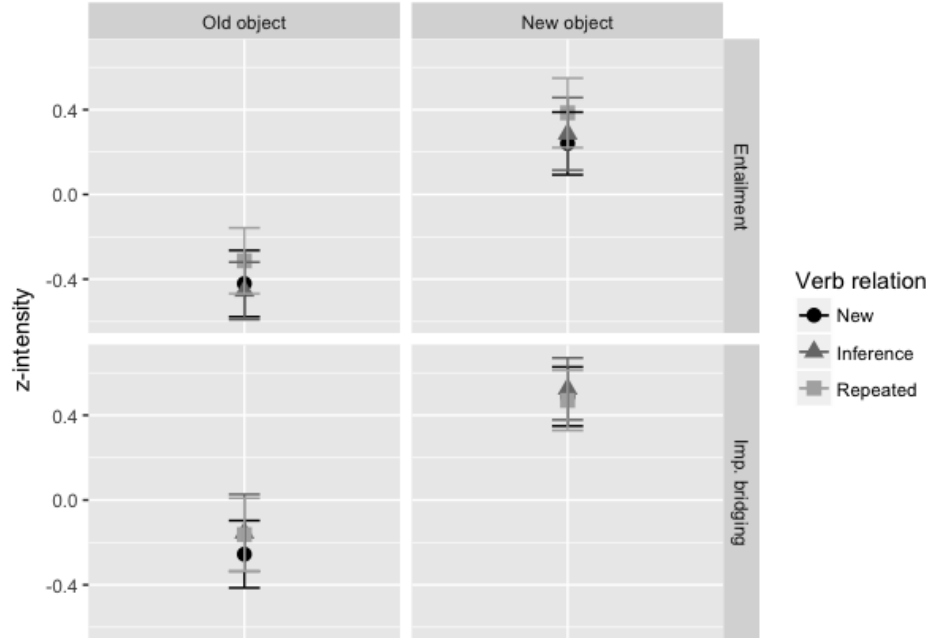


Figure 6.5: Experiment 9 results for z-transformed intensity, O1. Horizontal split: Object status. Vertical split: Inferencing relation (items 1-6 versus 7-12). Error bars: 95% confidence interval.

effects regression models were constructed for items 1-6 (with entailment as the inferencing relation) and items 7-12 (implicational bridging). Each model contained an interaction between the clause’s verb relation and its object relation as well as the component main effects, plus random intercepts for participant and item.¹

For z-transformed intensity on O1, among the entailment items, there was a significant main effect of object status ($p < .001$), while the main effect of verb status ($p > .1$) and the interaction ($p > .8$) were not significant. For the bridging items, the main effect of object was significant ($p < .001$), while the effect of verb status ($p > .6$) and the interaction ($p > .8$) were not significant.

For z-transformed f_0 on O1, in the entailment model, there were significant main effects of verb status ($p < .01$) and object status ($p < .001$), while the interaction was not significant ($p > .8$). To explore the significant effect of verb status, paired comparisons were carried out between the verb conditions averaged across both object types. The means with new and

1. `lmer` model specification: $\text{Response} \sim \text{Verb} * \text{Object} + (1 \mid \text{Participant}) + (1 \mid \text{Item})$.

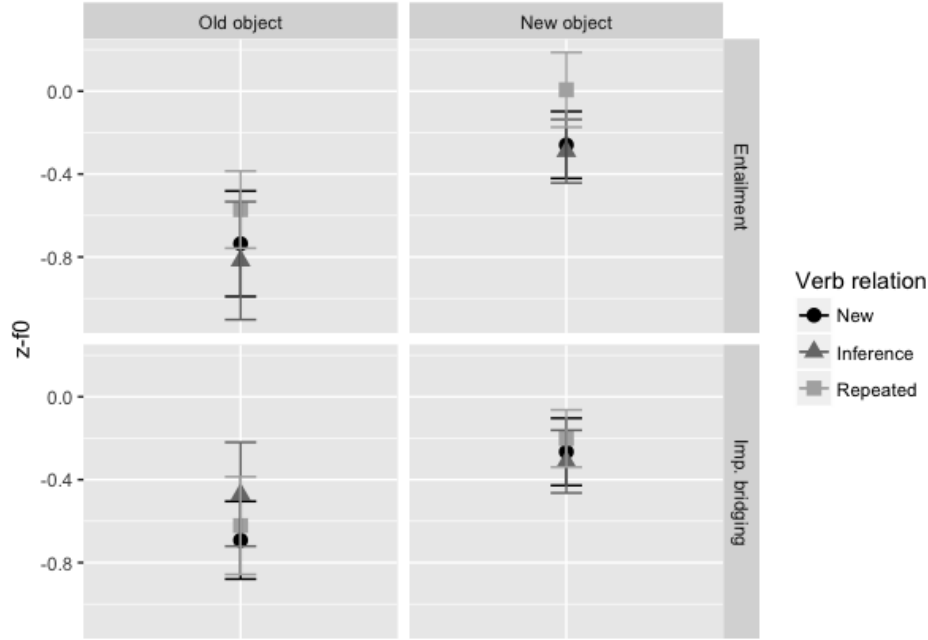


Figure 6.6: Experiment 9 results for z-transformed f_0 , O1. Horizontal split: Object status. Vertical split: Inferencing relation (items 1-6 versus 7-12). Error bars: 95% confidence interval.

repeated ($p < .05$) and inference and repeated ($p < .01$) verbs were significantly different, while the means with new and inference verbs were not significantly different ($p > .7$).

For the bridging items, there was a significant main effect of object relation ($p < .001$), while the effect of verb status ($p > .3$) and the interaction ($p > .2$) were not significant.

For duration on O1, for both sets of items, there was a significant main effect of object status (p 's $< .001$). The main effect of verb status (entailment $p > .3$, bridging $p > .2$) and the interaction (p 's $> .3$) were not significant.

The results for O2 are shown in Figures 6.8-6.10.

For z-transformed intensity on O2, there was a significant main effect of object relation for both the entailment items ($p < .001$) and the bridging items ($p < .05$). The main effect of verb relation (entailment $p > .1$, bridging $p > .4$) and the interaction (entailment $p > .9$, bridging $p > .3$) were not significant in either model.

For z-transformed f_0 on O2, none of the main effects or interactions were significant in either model (verb status, bridging $p > .1$; all other p 's $> .3$).

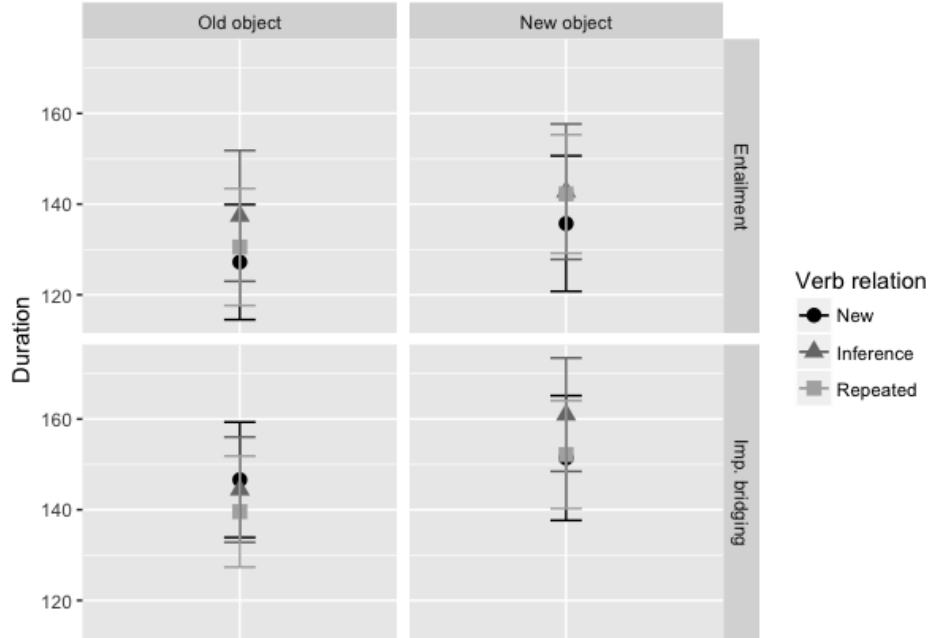


Figure 6.7: Experiment 9 results for duration, O1. Horizontal split: Object status. Vertical split: Inferencing relation (items 1-6 versus 7-12). Error bars: 95% confidence interval.

For duration on O2, among the entailment items, neither of the main effects (verb status $p > .8$, object status $p > .4$) nor the interaction ($p > .3$) were significant. For the bridging items, the main effects of verb status ($p > .3$) and object status ($p > .1$) were not significant, while there was a marginal interaction between the two ($p < .1$).

Verbs

The results for z-transformed intensity, z-transformed f_0 , and duration on V1 are shown in Figures 6.11-6.13.

For z-transformed intensity on V1, in the entailment model, neither of the main effects (verb status $p > .2$, object status $p > .6$) were significant, nor was the interaction ($p > .6$). In the bridging model, there was a marginal effect of verb status ($p < .1$), while the main effect of object status ($p > .6$) and the interaction ($p > .6$) were not significant.

For z-transformed f_0 on V1, in the entailment model, there were significant main effects of both verb status ($p < .001$) and object status ($p < .01$), while the interaction was not significant ($p > .4$). Paired comparisons were made between the different verb relations irrespective of

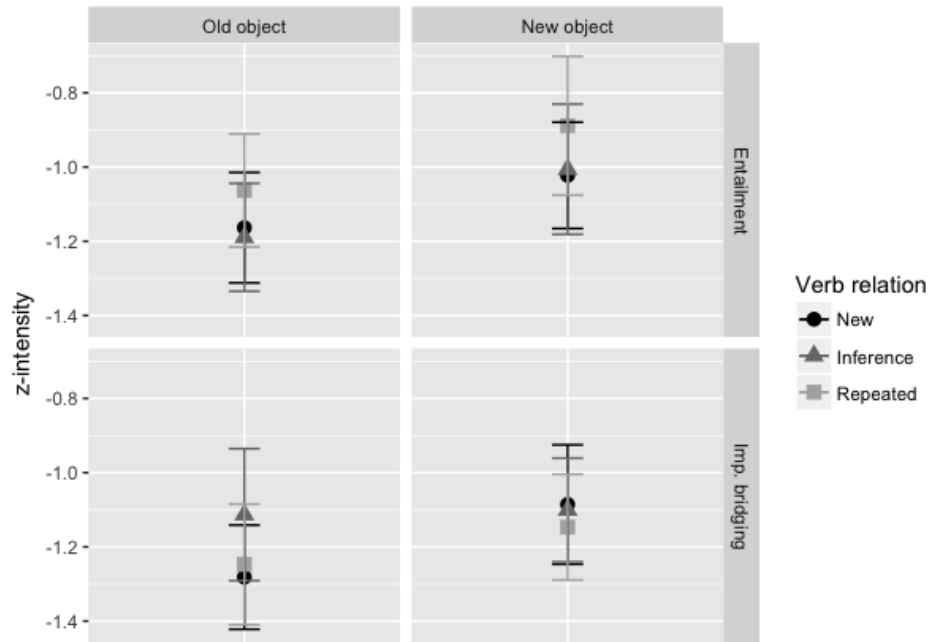


Figure 6.8: Experiment 9 results for z-transformed intensity, O2. Horizontal split: Object status. Vertical split: Inferencing relation (items 1-6 versus 7-12). Error bars: 95% confidence interval.

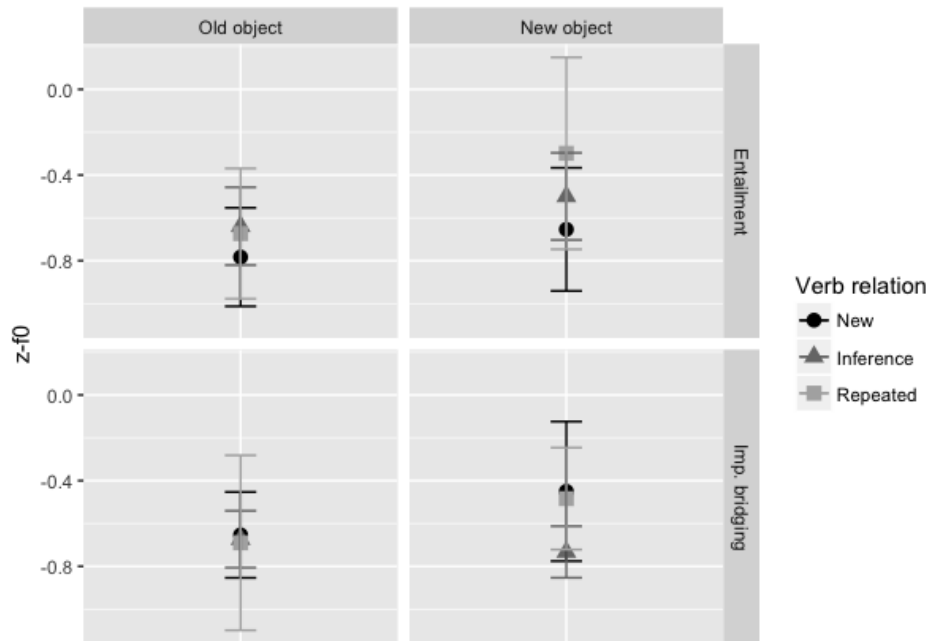


Figure 6.9: Experiment 9 results for z-transformed f_0 , O2. Horizontal split: Object status. Vertical split: Inferencing relation (items 1-6 versus 7-12). Error bars: 95% confidence interval.

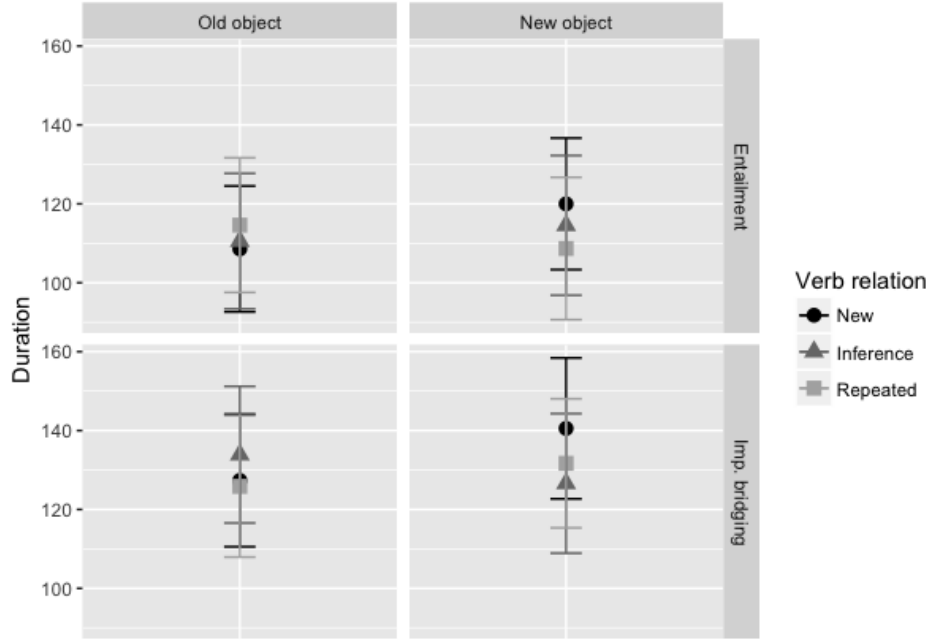


Figure 6.10: Experiment 9 results for duration, O2. Horizontal split: Object status. Vertical split: Inferencing relation (items 1-6 versus 7-12). Error bars: 95% confidence interval.

object status using estimated marginal means. The means for new and repeated ($p < .05$) and inference and repeated ($p < .001$) verbs were significantly different, while the means for new and inference verbs were not significantly different ($p > .5$).

For f_0 among the bridging items, there were significant main effects of verb status ($p < .05$) and object status ($p < .001$) as well as a significant interaction ($p < .05$). To explore the significant interaction, paired comparisons were carried out between conditions with the same object status but different verb relations. When the object was old, the means for new and repeated ($p < .01$) and inference and repeated ($p < .01$) verbs were significantly different, while the means for new and inference verbs were not significantly different ($p > .9$). When the object was new, the means were not significantly different for any pair of verb status conditions ($p > .9$).

For duration on V1, in the entailment model, there was a marginal effect of verb status ($p < .1$), while the main effect of object status ($p > .7$) and the interaction ($p > .6$) were not significant. In the bridging model, there was a significant main effect of object relation

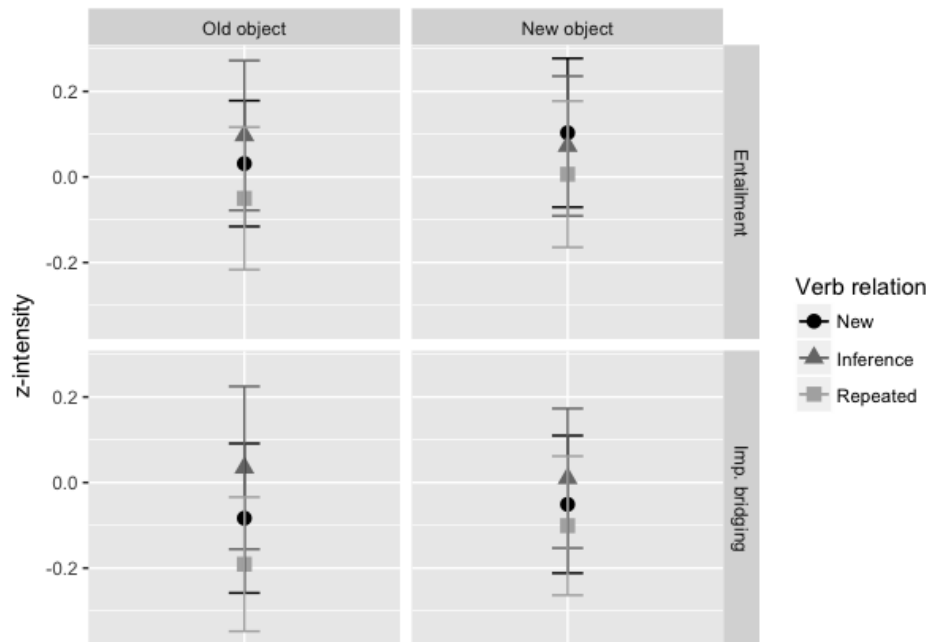


Figure 6.11: Experiment 9 results for z-transformed intensity, V1. Horizontal split: Object status. Vertical split: Inferencing relation (items 1-6 versus 7-12). Error bars: 95% confidence interval.

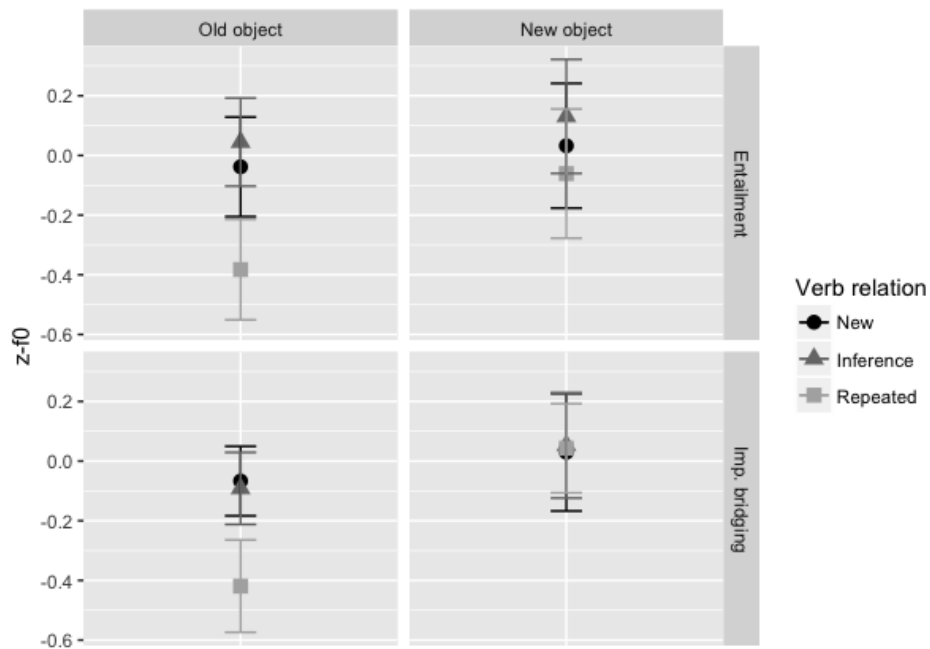


Figure 6.12: Experiment 9 results for z-transformed f_0 , V1. Horizontal split: Object status. Vertical split: Inferencing relation (items 1-6 versus 7-12). Error bars: 95% confidence interval.

($p < .05$), while the main effect of verb status ($p > .5$) and the interaction ($p > .3$) were not significant.

The results for V2 are shown in Figures 6.14-6.16.

For z-transformed intensity on V2, among the entailment items, the main effect of verb relation ($p < .001$) and the interaction ($p < .001$) were significant, while the main effect of object status was not ($p > .2$). Paired comparisons were made between verb conditions and within object conditions using estimated marginal means. With an old object, the means for new and repeated and for inference and repeated verbs were significantly different (p 's $< .001$), while the means for new and inference verbs were not significantly different ($p > .7$). With a new object, the means for new and repeated verbs ($p < .01$) and inference and repeated verbs ($p < .05$) were significantly different, but the means for new and inference verbs were not significantly different ($p > .7$).

In the bridging model, the main effect of verb status ($p < .001$) and the interaction ($p < .001$) were significant, while the main effect of object status was not significant ($p > .8$). Paired comparisons showed that with an old object, the means for new and repeated and for inference and repeated verbs were significantly different (p 's $< .001$), while the means for new and inference verbs were not significantly different ($p > .9$). With a new object, there was no significant difference in the means for any pair of verb conditions (new-repeated $p > .6$, inference-repeated $p > .7$, new-inference $p > .9$).

For z-transformed f_0 on V2, in the entailment model, there was a significant main effect of verb relation ($p < .001$) and a significant interaction ($p < .001$), while the main effect of object status was not significant ($p > .4$). Paired comparisons indicated that with an old object, the means for new and repeated and for inference and repeated verbs were significantly different (p 's $< .001$), whereas the means for new and inference verbs were not significantly different ($p > .7$). With a new object, the means were not significantly different for any pair of verb conditions (new-repeated $p > .4$, inference-repeated $p > .2$, new-inference $p > .9$).

For the bridging model, the main effects of verb status and object status and the inter-

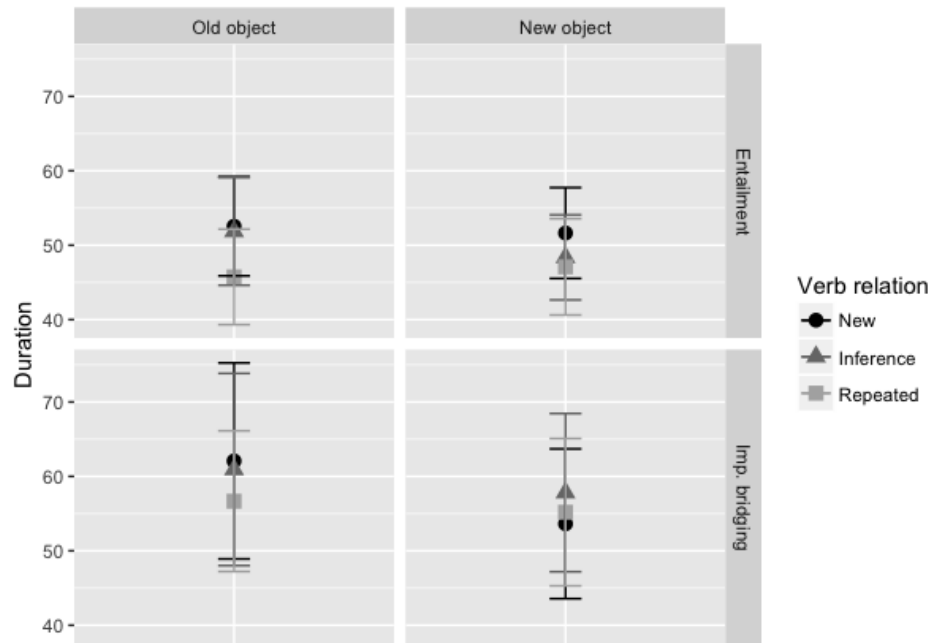


Figure 6.13: Experiment 9 results for duration, V1. Horizontal split: Object status. Vertical split: Inferencing relation (items 1-6 versus 7-12). Error bars: 95% confidence interval.

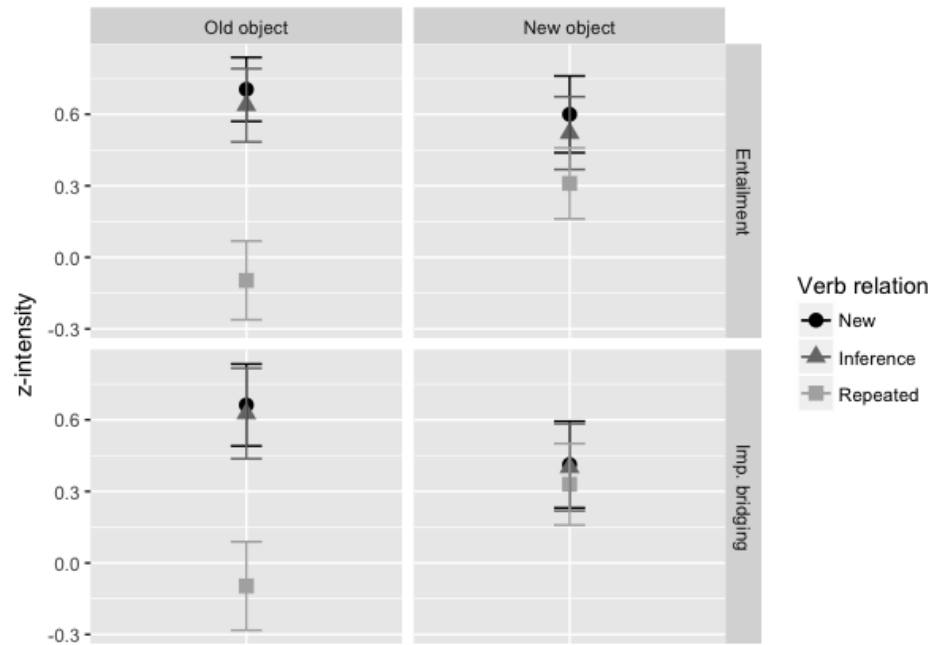


Figure 6.14: Experiment 9 results for z-transformed intensity, V2. Horizontal split: Object status. Vertical split: Inferencing relation (items 1-6 versus 7-12). Error bars: 95% confidence interval.

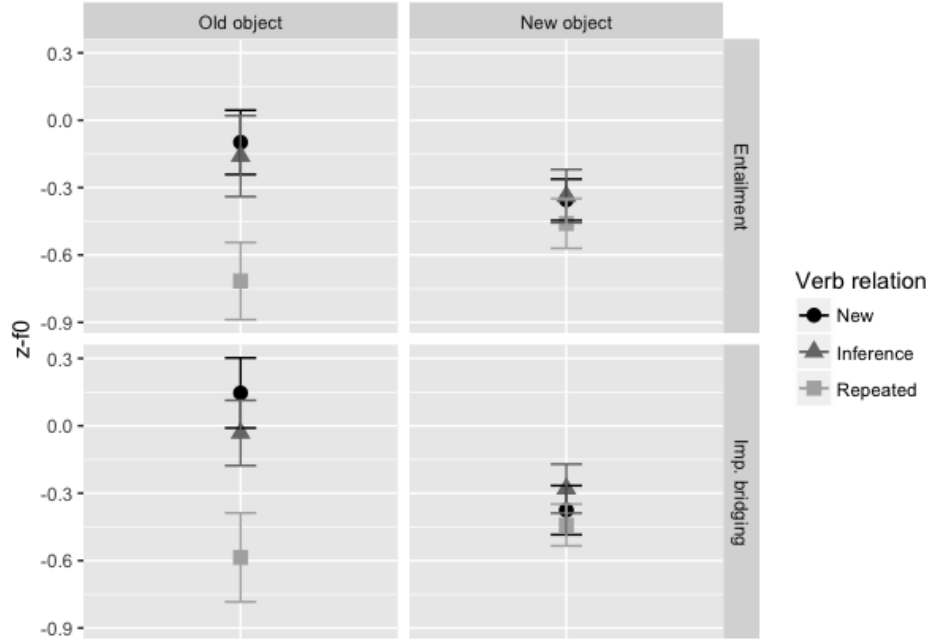


Figure 6.15: Experiment 9 results for z-transformed f_0 , V2. Horizontal split: Object status. Vertical split: Inferencing relation (items 1-6 versus 7-12). Error bars: 95% confidence interval.

action were all significant (p 's < .001). Paired comparisons indicated that with an old object, the means for new and repeated and for inference and repeated verbs were significantly different (p 's < .001), while the means for new and inference verbs were not significantly different (p > .1). With a new object, the means were not significantly different for any pair of verb conditions (new-repeated p > .6, inference-repeated p > .1, new-inference p > .6).

For duration on V2, the entailment model showed a significant main effect of verb status (p < .001), while the main effect of object status and the interaction were marginal (p 's < .1). Paired comparisons were carried out within object conditions. With an old object, the means for new and repeated verbs and for inference and repeated verbs were significantly different (p 's < .001), and the means for new and inference verbs were not significantly different (p > .8). With a new object, the means for new and repeated verbs were significantly different (p < .05), while the means for inference and repeated (p > .6) and for new and inference verbs (p > .1) were not significantly different.

For the bridging model, the main effects of verb status (p < .001) and object status

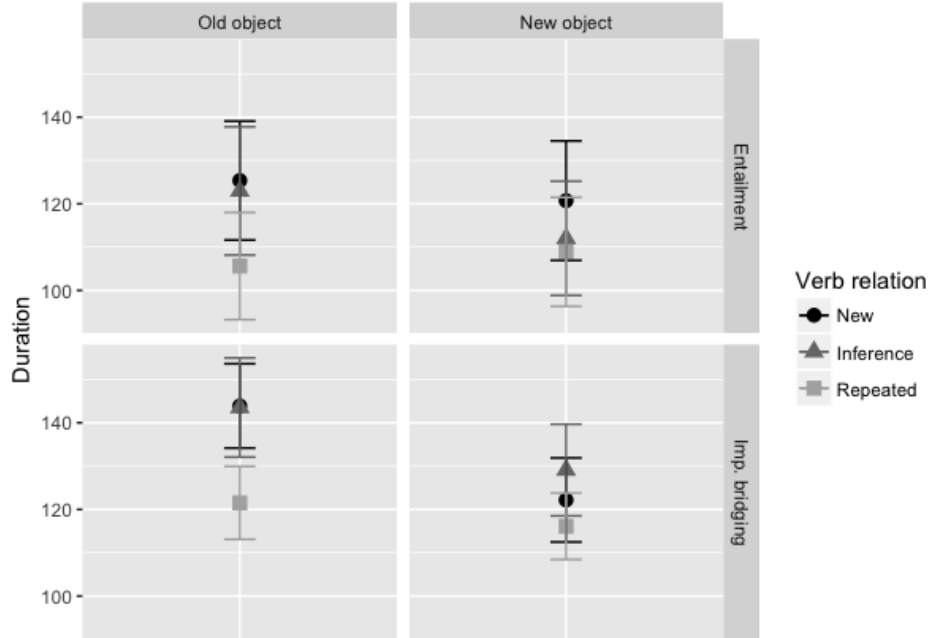


Figure 6.16: Experiment 9 results for duration, V2. Horizontal split: Object status. Vertical split: Inferencing relation (items 1-6 versus 7-12). Error bars: 95% confidence interval.

($p < .001$) and the interaction ($p < .05$) were all significant. Paired comparisons showed that with an old object, the means for new and repeated verbs and for inference and repeated verbs were significantly different (p 's $< .001$), while the means for new and inference verbs were not significantly different ($p > .9$). With a new object, the means for inference and repeated verbs were significantly different ($p < .01$), while the means for new and repeated ($p > .3$) and for new and inference verbs ($p > .2$) were not significantly different.

Subjects

The results for subject syllables are shown in Figures 6.17-6.19. For the entailment model for z-transformed f_0 on S, the main effect of verb status was marginal ($p < .1$). For all other models, neither of the main effects nor the interaction attained significance (all p 's $> .1$).

6.2.8 Discussion

As in the analysis subsection, the second-clause constituents are discussed in reverse order: objects, then verbs, then subjects.

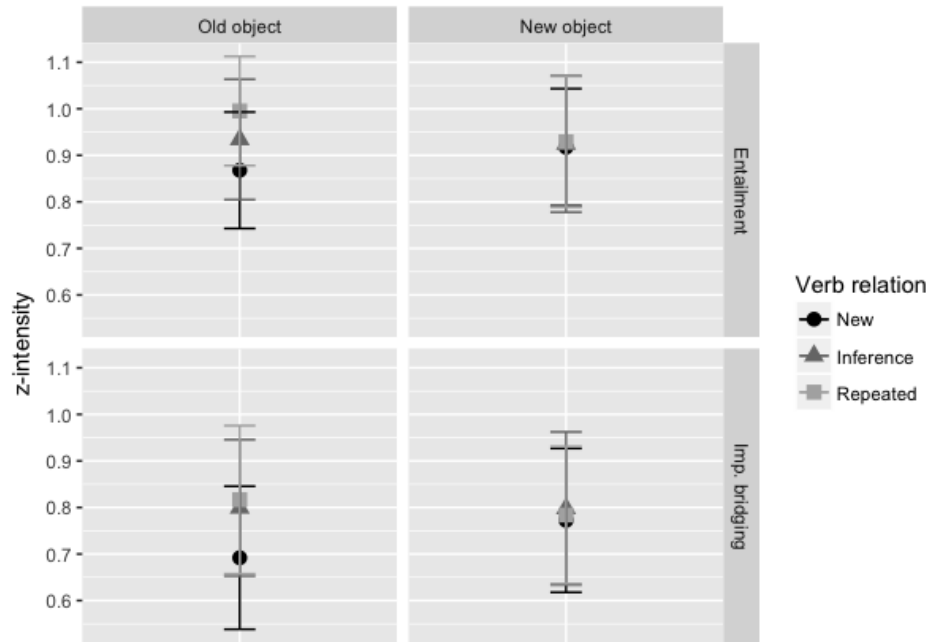


Figure 6.17: Experiment 9 results for z-transformed intensity, S. Horizontal position: Syllable within critical clause. Horizontal split: Object status. Vertical split: Inferencing relation (items 1-6 versus 7-12). Error bars: 95% confidence interval.

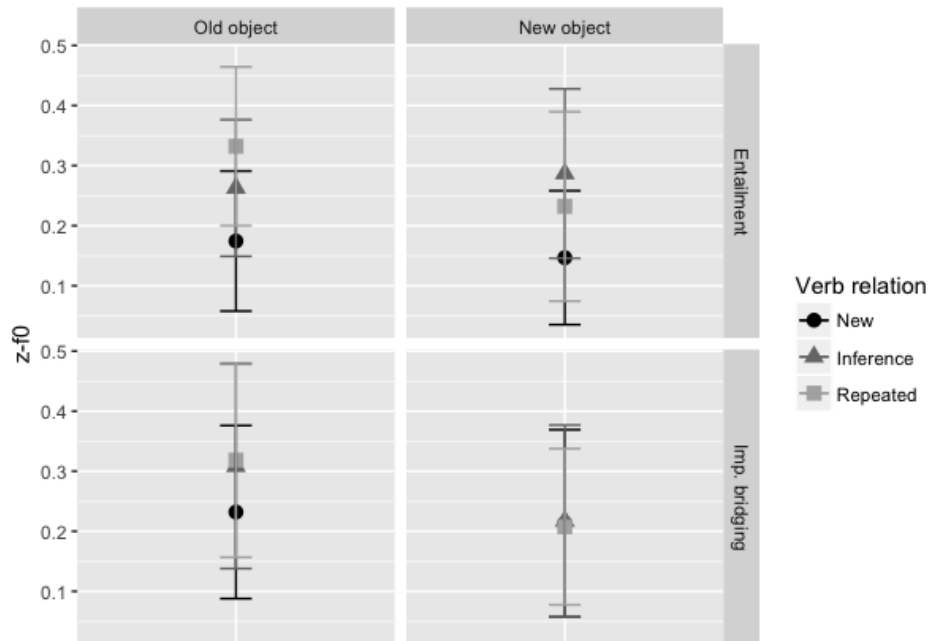


Figure 6.18: Experiment 9 results for z-transformed f_0 , S. Horizontal position: Syllable within critical clause. Horizontal split: Object status. Vertical split: Inferencing relation (items 1-6 versus 7-12). Error bars: 95% confidence interval.

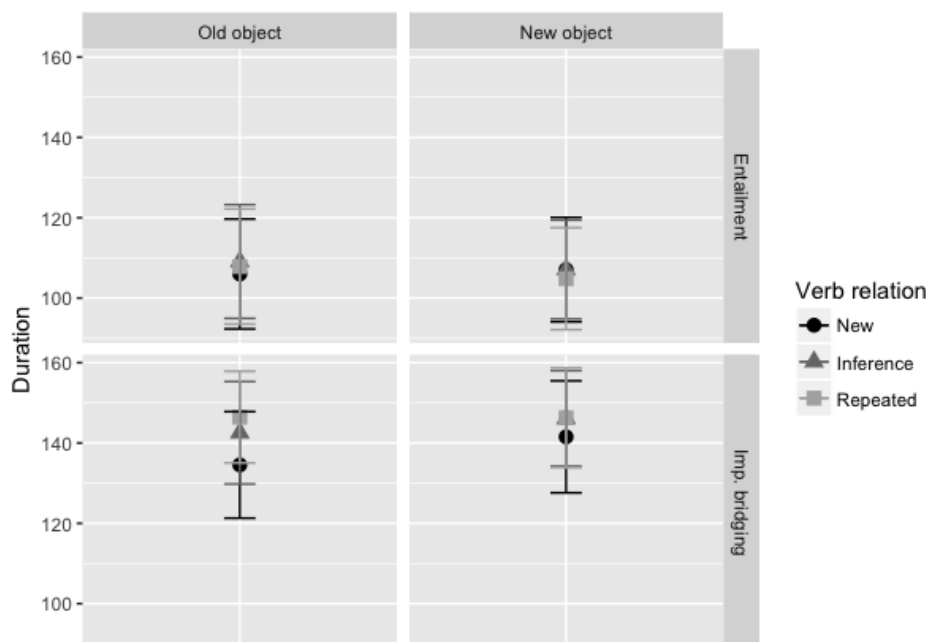


Figure 6.19: Experiment 9 results for duration, S. Horizontal position: Syllable within critical clause. Horizontal split: Object status. Vertical split: Inferencing relation (items 1-6 versus 7-12). Error bars: 95% confidence interval.

Objects

The results for the object syllables suggest that the experimental paradigm and analysis successfully elicited and detected differential emphasis behavior from the participants. The first syllable of each object was lexically stressed, meaning this syllable should carry the bulk of the information about the accent status of this word (Ladd, 1996). The main effect of object status was significant in all six of the models for this syllable, with inspection of the means indicating that the values for discourse-new objects were significantly higher than those for repeated objects. This is exactly the expected result; a discourse-new object is canonically expected to carry a pitch accent, which is associated with higher phonetic values, while a repeated object should be deaccented and have lower values.

Interestingly, for the entailment items only, there was a significant main effect of verb status in determining the z-transformed value of f_0 . Paired comparisons indicated that when the verb was repeated, the f_0 value on O1 was significantly higher than when the verb was new or entailed. This effect may have been driven by deaccenting of repeated verbs. Since

a repeated verb canonically should be deaccented, it is not necessarily surprising that an object that follows it should have higher phonetic values than normal; that is, an object in the context of a repeated verb might be produced with slightly more emphasis than an object in the context of a new (accented) verb. However, this trend did not generalize to f_0 for the bridging items or the other phonetic correlates.

For O2, there was a significant effect of object status on z-transformed intensity for both the entailment and the bridging items, whereas there were no significant effects for the other two phonetic correlates. Inspection of the means suggests that the intensity values for new objects were significantly higher than for old objects. This suggests that the second object syllable exhibited similar emphasis behavior to the first syllable, with new objects being more emphasized than old objects, but to a lesser degree, since the emphasis difference was only detected for one phonetic correlate. This finding is not highly surprising. Since the second object syllable was always unstressed, it can be expected to carry relatively little prosodic information. Additionally, since this syllable is unstressed and in sentence-final position, it is reasonable to expect that it would be highly reduced regardless of the full word's emphasis status. This is evidenced by the high number of f_0 values that were removed from analysis for this syllable due to creaky voice.

Verbs

With the experimental and analytical paradigm's sensitivity to prosodic emphasis verified, the analysis can turn to the main constituent of interest, verbs. The critical question of the experiment is whether inference verbs exhibit prosodic realizations similar to those of new verbs, suggesting that they were accented; similar to repeated verbs, suggesting that they were deaccented despite their surface nonidentity with the first-clause verb; or a pattern that is categorically separate from both new and repeated verbs.

V1 is the unstressed syllable, and thus should carry relatively less prosodic information than V2. There were some telling effects even on V1, however. For the z-transformed f_0 values, there was a reliable effect that the values for repeated verbs in the context of an old

objects were significantly lower than the values for either new or inference verbs in the same context. In addition, the values for verbs in the context of an old object across the board were lower than the values for verbs in the context of a new object.

For duration on V1, there was a marginal effect of verb relation for the entailment items, with paired comparisons suggesting that the values for repeated verbs were lower than for either new or inference verbs, which were not significantly different from one another. For the bridging items, there was a significant effect of object status, with the values in the context of a new object lower than those in the context of an old object.

The results for f_0 and for duration among the entailment items speak to the experiment's main question of whether inferable constituents are deaccented to a degree that distinguishes them from discourse-new verbs. The most apparent effect comes when the following object is discourse-old and canonically deaccented. In this case, the verb is in a position to potentially carry a nuclear pitch accent, since it is the last constituent before a deaccented constituent. In these conditions, the f_0 and bridging-item duration results suggest that new verbs were accented and repeated verbs deaccented, as canonically expected. However, the values for inference verbs were not reliably different from the values for new verbs in any of these conditions, suggesting that they were not deaccented.

When the object is new, it is canonically accented, and as the most embedded constituent in the clause, it carries a nuclear pitch accent in these conditions. Thus, it would not be surprising that emphasis differences on the verb might be neutralized in these cases, as at most it can carry a non-nuclear pitch accent. Indeed, in those analyses with significant main effects of verb status or significant interactions (f_0 ; bridging-item duration), paired comparisons showed no differences between the values for the different verb conditions in the context of a new object.

It is less clear what might be driving the effects of object status on the realization of V1. For the f_0 analysis, the values were reliably lower for a verb in the context of an old object than in the context of a new object. However, among the bridging items, the durations for V1

were higher in the context of an old object than a new object. These results are apparently contradictory, with the f_0 values indicating that V1 is less emphasized in the context of an old object and the duration values indicating that V1 is more emphasized in this context. It is not necessarily surprising that the discourse status of the object (or rather, its emphasis status) would affect the phonetic values on V1, as this syllable might receive more or less relative emphasis depending on the status of the object. However, it is difficult to reconcile these apparently competing observations, and this remains a topic for future consideration.

V2, as the stressed syllable of the verb, was expected to carry more reliable information about the prosodic status of the verb. Indeed, the results for this syllable show more agreement with one another, and together they indicate that inference verbs as a group were generally not pronounced differently than new verbs, while repeated verbs were – in other words, inferable verbs were not deaccented.

For each of the phonetic correlates and item groups, there was a significant interaction of verb status and object status in determining the phonetic correlate, with the exception that the interaction was only marginally significant for duration on the entailment items. These significant interactions indicate that the effect of the verb status manipulation on the phonetic values varies depending on the object relation in the sentence, so the different types of object relation will be discussed separately.

For sentences with a new object, there were only minor suggestions of an effect of verb status on the prosodic realization of the verb. For z-transformed intensity, there were significant differences between new and repeated and between inference and repeated verbs among the entailment items, while there were no significant differences by verb status for the bridging items. For z-transformed f_0 , there were no differences by verb status on V2 for either of the groups of items. The duration model for the entailment items showed a significant difference between new and repeated verbs, while the model for bridging items showed a significant difference between inference and repeated verbs.

Considered together, the results for verbs produced in the context of a new object indicate

that there was a slight trend toward repeated verbs being pronounced with less prominence than new and inference verbs. Repeated verbs had significantly lower z-intensity than both new and inference verbs among the entailment items, and also had lower duration than new verbs for the entailment items and than inference verbs among the bridging items. However, this trend was far from consistent for all combinations of phonetic correlate and item group, making it difficult to characterize conclusively that the verb status manipulation affected the prosodic realization of the verb in the context of a new object.

As discussed above for V1, it is not necessarily surprising that there was only a weak effect of the verb relation manipulation in the context of a new object. A new object in this experimental paradigm should canonically carry a nuclear pitch accent, and there were no *a priori* expectations regarding how readily phonetic differences in the realization of verbs in a prenuclear position would be detected. However, it is worth noting that there were no parts of the analysis for which a significant difference between the phonetic correlates for new and inferable verbs were found (the closest being the duration comparison for entailment items).

In contrast to the new-object conditions, the old-object conditions show a clear effect that repeated verbs were deaccented to the exclusion of both new verbs and inference verbs. Paired comparisons for each combination of phonetic correlate and item group showed that the phonetic values for repeated verbs were significantly different from those for both new and inference verbs, which in turn were never significantly different from one another.

The clearer trend on verbs produced in the context of an old object is likely due to the canonical deaccenting of these objects. As a result of this, any pitch accent on the verb will be a nuclear pitch accent. The relatively larger range of phonetic values observed on old-object verbs compared to new-object verbs is consistent with observations in the literature that nuclear pitch accents exhibit a tighter connection to information status than do prenuclear accents (Chodroff and Cole, 2018).

Subjects

There were no significant effects of either the verb manipulation or the object manipula-

tion on the phonetic correlates of emphasis for the nuclei of the second-clause subjects. The subject data did show some intriguing numerical trends; for instance, among the bridging items with an old object, subjects in the context of a new verb had apparently lower values than subjects in the context of inferable or repeated verbs. This may be due to a slight trend toward lower emphasis on a subject when it is followed by an accented verb, although this proposal is difficult to reconcile with the lack of such a trend on subjects adjacent to inferable verbs, which the verb analysis indicates were also accented.

Despite the presence of such trends, however, the range of phonetic values observed for subjects is generally compressed compared to those observed on other syllables, and no trends attained significance. All subjects in this experiment were discourse-new, and as such were all expected to bear a pitch accent. While it was expected that certain subjects would bear a nuclear pitch accent (when followed by a deaccented verb and object) and others would bear a prenuclear accent (when followed by an accented verb and a deaccented object), it appears that no phonetic difference in the realizations of these two accents on the subject syllable was detected by the current analytical paradigm.

Summary and implications

To summarize the findings for each of the second-clause syllables, there were two major significant results observed on objects and verbs, while there were no significant findings on the subject syllable. First, the results for O1 and, to a lesser extent, O2 indicated that new objects had significantly higher phonetic correlates than old objects. This finding was expected and indicates that the experimental and analytical paradigm successfully elicited and detected differential emphasis behavior from the participants.

Second, the results from V1 and particularly V2 showed that when a verb was in a position to potentially receive a nuclear pitch accent (i.e., was followed by a deaccented object), repeated verbs had significantly lower phonetic correlates than new and inferable verbs, whose phonetic correlates were not significantly different from one another. New and repeated verbs were expected to be canonically accented and deaccented, respectively, and

the phonetic values observed for these two verb types in the context of an old object are consistent with this expectation. Thus, the fact that the phonetic values for inference verbs were significantly different from those for repeated verbs and not from those for new verbs suggests that inference verbs were generally accented in this position.

In contrast, there were no significant differences in the phonetic realization of verbs by discourse status when the verb was not in a position to receive a nuclear pitch accent (i.e., when followed by a new object, which carried a nuclear accent). This is not necessarily surprising; the realization of prenuclear pitch accents has received less attention than that of nuclear pitch accents, but prenuclear accents have been noted as perceptually weaker than nuclear accents. The fact that new and repeated verbs did not show any phonetic differences indicates that the experimental paradigm could not detect differences in phonetic realization between canonically accented and deaccented verbs in prenuclear position, if any exist.

Before discussing the implications of these findings for the theory of emphasis licensing, it is worth considering one possible issue with the experiment. Although certain phonetic correlates of prosodic emphasis have been identified in the literature, it has also been noted that the impressionistic perception of emphasis by native speakers is sometimes independent of the apparent phonetic reality of the speech signal. Although the results discussed above appear to indicate that inference verbs had a similar prosodic status to new verbs, *contra* the empirical claims of the bulk of the theoretical literature on emphasis licensing, it is possible that analysis of the acoustic signal alone is not a sufficient characterization of the phonological emphasis status of the recorded constituents. To address this possibility, Experiment 10, presented below, was carried out to elicit native listeners' perception of the emphasis status of the constituents produced in Experiment 9.

6.3 Experiment 10: Perception of emphasis in production data

Experiment 9 was a production study that recorded speakers' productions of verbs with varying discourse status: new, inferable, or repeated. The analysis focused on three phonetic

correlates of emphasis on the verbs' nuclei – mean intensity, mean f_0 , and duration. The results indicated that in a position where the verb could possibly bear a nuclear pitch accent, the phonetic realizations of new and inference verbs were categorically different from those of repeated verbs, suggesting that new and inference verbs were both accented, while repeated verbs were deaccented.

These results are potentially problematic for several accounts of emphasis licensing presented in the previous chapter, which predict that deaccenting should be highly likely or even mandatory on constituents that are readily inferable from a linguistic antecedent. One possibility that remains open after the phonetic analysis, however, is that the three phonetic correlates analyzed in Experiment 9 do not fully capture the impressionistic reality of emphasis on the recorded constituents as perceived by native English listeners. In other words, although the three analyzed correlates suggest that inference verbs were produced similarly to new verbs, i.e., accented, there may be other factors influencing listeners' phonological categorization of these constituents such that they sound deaccented. If this is the case, the results of Experiment 9 are less problematic for the theories of emphasis licensing above, although the nature of deaccenting of identical versus inferable material would be phonetically different in kind.

To explore this possibility, Experiment 10 was carried out to elicit native listeners' judgments of the emphasis status of the constituents recorded in Experiment 9. In this experiment, participants listened to the second SVO clause recordings produced in Experiment 9 isolated from the first clause to obscure the discourse status of the constituents. On each trial, they rated either the verb or object as “emphasized” or “not emphasized”. The goal of the experiment is to compare the proportion of times constituents for each level of discourse status are rated as emphasized to determine whether the conclusions of the phonetic analysis sufficiently describe the emphasis status of the constituents or whether there are factors other than those analyzed in Experiment 9 that affect the perception of emphasis.

6.3.1 Design and materials

The materials for the experiment were recordings of the production trials made during Experiment 9. Recall that each production trial consisted of three sentences, of which the critical middle sentence was of the form *SVO and SVO*. For Experiment 10, the second *SVO* clause was clipped from each Experiment 9 trial recording and used as a trial stimulus. This allowed for participants to rate the emphasis status of the critical second-clause stimuli without any knowledge regarding the discourse status of the constituent – for instance, participants must determine whether a particular object is emphasized without having heard the new or old object that was produced in the first clause of the production trial and which determines the discourse status of the critical object.

Participants in Experiment 9 reliably produced a pause after the end of the critical second clause, which facilitated the placement of the ending point of the clipped recordings. In general, due to the Experiment 9 instructions designed to elicit clear speech, production participants often also produced a pause between the word *and* and the subject of the second clause. When there was not a pause after the word *and*, the starting point of the clipped recording was placed with the help of visual inspection of the recording's spectrogram in Praat to avoid inclusion of the final [n] or [d] from the word *and*.

Like Experiment 9, Experiment 10 featured a 2 (object status) x 3 (verb status) design. Although the stimuli in Experiment 10 are all string identical, they were produced in 6 different combinations of discourse status in Experiment 9: the object could be new or old, and the verb could be new, inferable, or repeated. Table 6.4 shows example stimuli used for the experiment, including the first clause that preceded the production of each second clause and conditioned its constituents' discourse status, but was not actually played to participants in Experiment 10. In any given trial, participants were asked to rate the emphasis status of either the verb or the object.

Verb status	Object status	Hidden production context	Stimulus
New	New	<i>Sebastian punished Justin, and</i>	Nan surprised Amy.
New	Old	<i>Gabriel punished Amy, and</i>	Nan surprised Amy.
Inference	New	<i>Bradley astounded Justin, and</i>	Nan surprised Amy.
Inference	Old	<i>Ethan astounded Amy, and</i>	Nan surprised Amy.
Repeated	New	<i>Julian surprised Justin, and</i>	Nan surprised Amy.
Repeated	Old	<i>Benjamin surprised Amy, and</i>	Nan surprised Amy.

Table 6.4: Sample experiment 10 stimuli

6.3.2 Procedure

The experiment was conducted using Amazon Mechanical Turk and the same recruitment protocol as Experiment 8.

After providing informed consent and completing the demographic survey, participants read the instructions for the experiment. In this phase, participants were familiarized with the process of completing an experimental trial and were informed that their speakers or headphones should be connected, as the experiment involved playing sounds. After reading the instructions, participants completed two practice trials to ensure that they understood how the experiment would work and how to respond to the stimulus.

On each trial, participants were shown a “warning screen” informing them of the sentence they were about to hear. On this screen, they were shown one word of the sentence that they should focus on. A sample warning screen is shown at the top of Figure 6.20.

After viewing the “warning screen,” participants pressed any key on their keyboard to advance to the next screen. As soon as this screen loaded, the recorded Experiment 9 clause associated with the trial played automatically over the participant’s sound system. After listening to the sound file, participants used number keys on their keyboard or clicked the text indicating their preferred response to indicate whether they thought the target word was “emphasized” or “not emphasized.” A sample second screen is shown at the bottom of Figure 6.20.

Each of the 10 Experiment 9 participants read 12 items in 6 condition combinations, so there were a total of 720 clause recordings. In addition, since each individual trial in

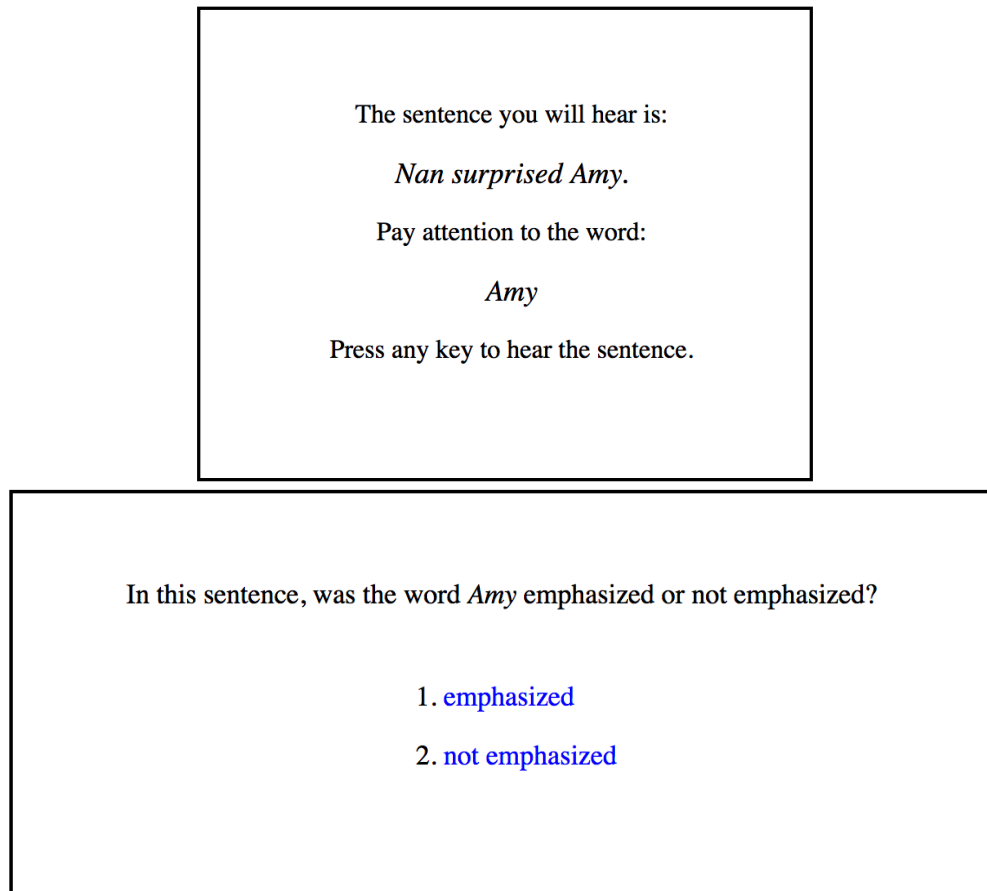


Figure 6.20: Sample screens from Experiment 10. The top screen displayed first. The second screen appeared and the recorded sentence played automatically when the participant pressed a key on the first screen.

Experiment 10 targeted either the verb or the object, there were 1440 unique trials in this experiment.

The trials were organized into 40 lists of 36 critical items each. Participants were assigned to lists in sequential order of beginning the experiment. For each participant, the order of presentation of the items in the list was randomized. Within a list, the included items cycled through the different possible values for speaker voice (Experiment 9 participant), item, verb status, and object status, so each list contained approximately equal number of trials for each level of each of these factors.

In addition to the 36 critical items, each list contained 4 additional filler items that were the same for every list. These items were intended to identify participants who were

not actually attending to the experiment. They followed the same general format as the critical items (e.g., rate *John* in *John greeted Rachel*), but the sound files were specifically constructed to have a correct answer as to whether the target word was emphasized or not emphasized.

6.3.3 *Participants*

The intended number of participants was 200, so 5 subjects would be assigned to each of the 40 lists and each trial would be completed an equal number of times. Due to an error, the actual number of participants was 201. Of these participants, 70 were female and 131 were male. The mean age of the participants was 33.7 years.

The data from 10 participants was excluded from analysis because they failed to self-report as native speakers of English in the demographic survey. Data from a further 8 participants was excluded because their mean reaction time across all trials was under 500 ms, indicating that they were entering responses without listening to the full sound recording.

Finally, data from a further 8 participants was excluded because they provided an unexpected response on more than 50%, i.e., at least three of four, of the constructed filler trials. A low threshold of concordance to the expected response on the fillers was chosen for inclusion because there was a large degree of variability in the responses on the fillers, with about 23% of responses in total giving the unexpected response. A total of 13 participants met the criterion for exclusion, but 3 had already been excluded as non-native speakers and 2 had been excluded for fast mean reaction times.

The mean time to complete the practice items, critical items, and fillers among those participants whose data was analyzed was 6 minutes, 32 seconds. Participants who completed the experiment were compensated with USD 1.00.

6.3.4 *Predictions*

Recall that the phonetic analysis of mean intensity, mean f_0 , and duration in Experiment 9 showed two major findings. First, the phonetic correlates of emphasis showed significantly higher values on new objects than on old objects, which was expected based on their discourse status. Second, on verbs uttered in the context of an old object, the phonetic correlates were significantly lower for repeated verbs than for new and inference verbs, suggesting that repeated verbs were deaccented while new and inference verbs were accented. By contrast, there were no significant effects of the verb status manipulation on verbs uttered in the context of a new object (i.e., in prenuclear position) or on any objects.

The goal of Experiment 10 is to use phonological perception of emphasis to verify whether the phonetic analysis presented in Experiment 9 is an adequate description of the phonological emphasis status of the constituents under investigation. If the phonetic description is adequate, then the perception results in Experiment 10 should be qualitatively similar to the findings from Experiment 9.

By contrast, it is also possible that the phonetic correlates of emphasis analyzed in Experiment 9 do not fully capture the phonological facts of emphasis. If this is the case, then there might be qualitative differences in listeners' ratings of emphasis by condition compared to the phonetic analysis presented previously. In particular, the perception results might indicate that inferable verbs were actually not emphasized, despite the fact that the phonetic analysis indicated that they were pronounced similarly to new verbs. This would support those accounts that predict that inferable verbs should be deaccented, which have difficulty accounting for the apparent observation from Experiment 9 that inferable verbs were accented.

6.3.5 *Results*

In addition to the participants who were excluded from analysis, data corresponding to the recordings from Experiment 9 trials with major speech errors were excluded from analysis.

A list of these trials is shown in Appendix K. This included 12 of 720 production trials for a total of 1.7% of trials.

The results for the perception of emphasis on objects, expressed as the proportion of trials with an “emphasized” rating by condition, are shown in Figure 6.21.

A visual inspection of the plot in Figure 6.21 suggests that participants’ perception of emphasis on the recorded objects was largely consistent with the Experiment 9 phonetic analysis. New objects appear to have been perceived as emphasized reliably more often than old objects. There are no clear differences in the perception of emphasis according to the discourse status of the verb, although in the entailment items it appears repeated verbs may have been perceived as less emphasized before an old object and more emphasized before a new object.

The results for the perception of emphasis on verbs are shown in Figure 6.22.

A visual inspection of the plot in Figure 6.22 likewise suggests results largely consistent with Experiment 9. In the context of an old object, participants appear to have perceived new and inference verbs as emphasized in comparable proportions of trials, whereas repeated verbs appear to have been rated as emphasized substantially less often. In the context of a new object, however, there do not appear to be any reliable differences in the proportion of times the verb was rated as emphasized for the different verb discourse status levels.

6.3.6 *Analysis*

Since the response data in Experiment 10 were binary, the results were analyzed using logistic mixed-effects regression. Separate models were constructed for the object responses and the verb responses, and within these two categories, for the items with entailment as the inferencing relation (items 1-6) and the items with implicational bridging as the inferencing relation (items 7-12). As in Experiment 9, the analysis of the results for objects are discussed first below, followed by the analysis of the verb responses.

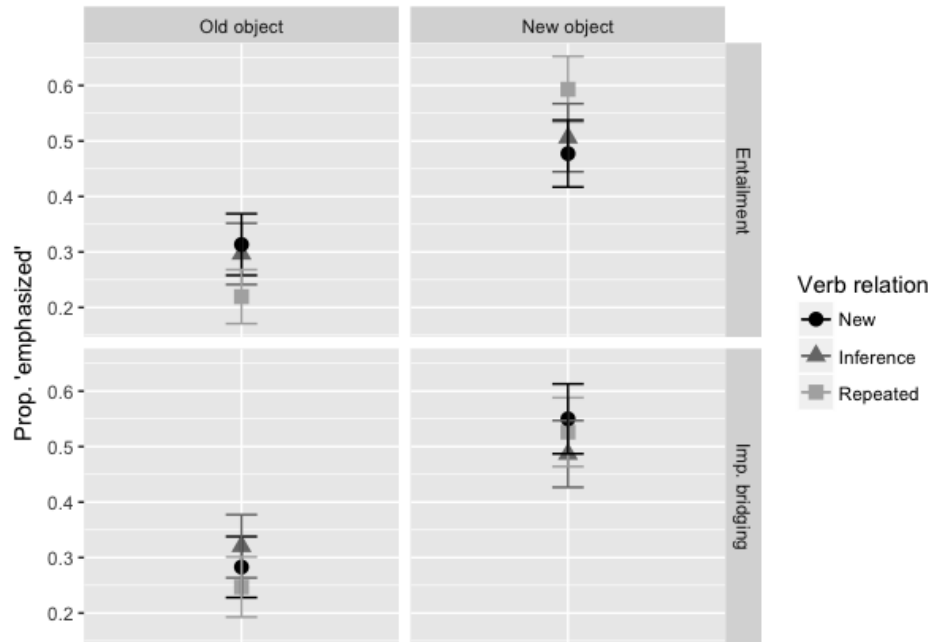


Figure 6.21: Experiment 10 results for object trials. Horizontal split: Object status. Vertical split: Inferencing relation (items 1-6 versus 7-12). Error bars: 95% confidence interval.

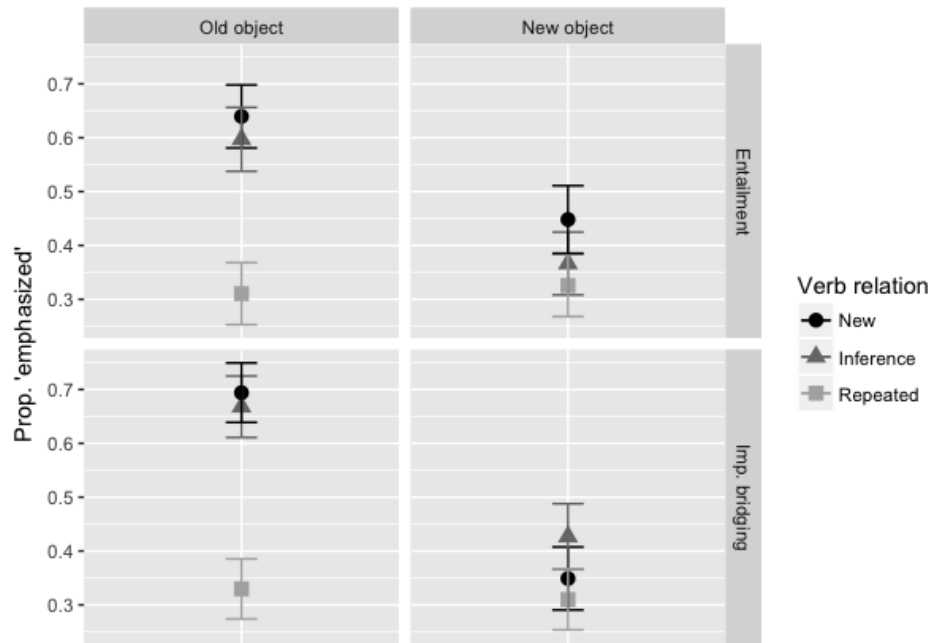


Figure 6.22: Experiment 10 results for verb trials. Horizontal split: Object status. Vertical split: Inferencing relation (items 1-6 versus 7-12). Error bars: 95% confidence interval.

Objects

For both the entailment items and the bridging items, two logistic MER models were compared to determine whether there was a significant interaction between verb status and object status in determining the proportion of emphasized ratings. The first model featured an interaction between the verb manipulation and the object manipulation plus the two component effects, as well as random intercepts for the voice heard in the recording (Experiment 9 participant), the participant rating the recording (Experiment 10 participant), and item.² The second model was the same, except that it featured only the component main effects, but no interaction.³

For the entailment items, model comparison indicated that the model with the interaction term exhibited a significantly better fit for the data ($p < .001$). Paired comparisons using estimated marginal means were carried out within object status levels and between verb status levels to explore the effect of verb status on emphasis ratings. When the object was old, the proportions of ratings for objects uttered in the context of a new versus inference verb were not significantly different ($p > .9$). By contrast, there were marginal differences in the proportions for old objects uttered in the presence of new versus repeated ($p < .1$) and inference versus repeated ($p < .1$) verbs.

When the object was new, the proportions for objects in the context of new and inference verbs were not significantly different ($p > .7$). However, the proportions for objects in the context of new and repeated verbs ($p < .01$) and in the context of inference and repeated verbs ($p < .05$) were significantly different.

In contrast to the entailment items, model comparison for the bridging items indicated that the model with an interaction between verb status and object status did not provide a significant better fit than the model with no interaction ($p > .1$). Further model comparison indicated that the model with main effects of verb status and object status (but no inter-

2. **glmer** model specification: $\text{Response} \sim \text{Verb} * \text{Object} + (1 \mid \text{Voice}) + (1 \mid \text{Participant}) + (1 \mid \text{Item})$.

3. **glmer** model specification: $\text{Response} \sim \text{Verb} + \text{Object} + (1 \mid \text{Voice}) + (1 \mid \text{Participant}) + (1 \mid \text{Item})$.

action) did not provide a significantly better fit than a model where the only main effect of object status ($p > .5$, with the same random effects), but did provide a better fit than a model with only a main effect of verb status ($p < .001$). The object-only model in turn provided a significantly better fit than a model with only the random effects ($p < .001$), indicating that only the object was a reliable predictor of the proportion of “emphasized” responses on the object in bridging items.

Verbs

For both the entailment and the bridging items, model comparison indicated that logistic MER models containing an interaction between verb status and object status provided significantly better fits for the response data than models containing only the component effects of this interaction (p 's $< .001$). This significant interaction indicates that the discourse status of verbs played a significant role in determining the proportion of trials on which they were perceived as emphasized, but this effect depended on whether the object of the sentence was discourse-old or discourse-new.

Paired comparisons were carried out using estimated marginal means within object status levels and between verb status levels. For the entailment items, with an old object, the proportions of emphasized responses were not significantly different for new or inference verbs ($p > .4$). The proportions were significantly different for new and repeated ($p < .001$) and for inference and repeated verbs ($p < .001$). When the object was new, the proportions were not significantly different for new versus inference ($p > .1$) or inference versus repeated verbs ($p > .4$), but the proportions for new and repeated verbs were significantly different ($p < .01$).

For the bridging items, when the object was old, the proportions of emphasized responses were not significantly different for new and inference verbs ($p > .8$). The proportions were significantly different for new and repeated ($p < .001$) and for inference and repeated verbs ($p < .001$). When the object was new, the proportions were not significantly different for new and repeated verbs ($p > .8$). The proportions for inference and repeated verbs were

significantly different ($p < .05$), while there was a marginal difference in proportions for new and inference verbs ($p < .1$).

6.3.7 Discussion

The significant role of object status in determining the emphasis proportions on objects among the entailment items indicates that new objects were perceived as emphasized significantly more often than old objects, which is consistent with the Experiment 9 phonetic analysis. Interestingly, verb status did play a role in determining the perception of emphasis among entailment-item objects. Old objects uttered adjacent to a repeated (i.e., deaccented) verb were perceived as marginally less emphasized than old objects adjacent to new and inference (by hypothesis, accented) verbs. Conversely, new objects uttered next to repeated verbs were perceived as more emphasized than new objects next to new and repeated verbs. These findings may point to general adjacency effects on the perception of emphasis: deaccented constituents (here, objects) that follow other deaccented constituents (verbs) may be perceived as less emphasized than constituents with comparable phonetic correlates that follow accented constituents. Likewise, accented constituents that follow a deaccented constituent may be perceived as more emphasized than constituents with similar correlates that follow accented constituents.

For the bridging items, objects were rated as emphasized significantly more often when they were new than when they were old. The verb status manipulation did not play a reliable role in determining perceptions of emphasis on the object for these items. This finding is consistent with the phonetic analysis from Experiment 9, which indicated that new objects were accented and old objects were deaccented, generally without reference to the discourse status of the verb that preceded the object.

The results for verbs likewise indicate similar results to the phonetic findings from Experiment 9. For both entailment and bridging items, when the object was old, meaning the verb was likely in a position to receive a nuclear pitch accent, repeated verbs were perceived

as emphasized significantly less often than new and inference verbs, which were perceived as emphasized a comparable proportion of times.

As in Experiment 9, the findings for verbs uttered in the context of a discourse-new object show a substantially less reliable effect of the verb status manipulation. Interestingly, there were some significant differences in the proportion of “emphasized” responses among verbs in new-object conditions. New verbs were rated as emphasized more often than repeated verbs among the entailment items, and inference verbs were rated as emphasized more often than repeated verbs among the bridging items. These findings recall some trends from the production data in Experiment 9; for instance, the durations of new-object discourse-new V2 nuclei were significantly higher than for new-object repeated V2 nuclei among the entailment items, whereas the durations of new-object inferable V2 nuclei were significantly higher than the corresponding repeated-verb durations among bridging items.

In sum, Experiment 10 suggests that the phonetic analysis from Experiment 9 faithfully represents the emphasis status of the constituents under investigation. The major findings of Experiment 9 were replicated in naive listeners’ categorical perception of emphasis in Experiment 10. First, new objects were rated as emphasized significantly more often than old objects, which aligns with the canonical expectation that new objects should be accented and old objects should be deaccented. Second, on old-object trials, repeated verbs were perceived as emphasized significantly less often than both new and inference verbs, which were rated as emphasized on approximately equal proportions of trials. This suggests that repeated verbs were deaccented in the production study, while new and inference verbs were generally accented.

Finally, the findings on verbs preceding a new object point to further concordance between the phonetic results and the perception results. New verbs were perceived as emphasized more often than repeated verbs in the entailment items, and inference verbs were perceived as emphasized more often than repeated verbs in the bridging items, echoing the findings of the analysis of duration on V2 in Experiment 9. This finding was far from robust in

the phonetic analysis, as it was not reflected in the z-transformed intensity or f_0 results. However, the agreement between the duration analysis and the perception results indicates that duration may be an important phonetic cue that predicts native listeners' perception of prosodic emphasis, and points to a minor trend toward repeated verbs being less emphasized than new and inference verbs when they precede a new object (i.e., when the verb is in prenuclear position).

6.4 General discussion

In summary, the results of Experiments 9 and 10 failed to detect evidence that speakers deaccented inferable verbs in production. Both experiments successfully detected differences in prosodic emphasis, as new objects were found to have higher values for the phonetic correlates of emphasis than old objects in Experiment 9 and were judged as more likely to be emphasized in Experiment 10. In old-object conditions, where the critical verb was in a position to potentially receive a nuclear pitch accent, there was a reliable phonetic and perceptual trend indicating that repeated verbs were deaccented, while new verbs were accented. In new-object conditions, there was a slight trend toward repeated verbs being less accented, with the duration and perceptual analysis indicating lower emphasis on repeated verbs than new verbs among the entailment items, and on repeated verbs than inference verbs among the implicational bridging items.

Turning to the critical inferable conditions, the experimental results agreed across the board that inferable verbs were not deaccented in nuclear position (before a deaccented old object). Inferable verbs had comparable phonetic correlates to new verbs, and much higher correlates than repeated verbs, in this position. Listeners also rated inferable verbs and new verbs as emphasized a comparable proportion of the time, while repeated verbs were rated as emphasized less often. In prenuclear position (before an accented new object), the results, unsurprisingly, were less clear. However, there were instances, such as the duration and perceptual analysis, where inference verbs (but not new verbs) were analyzed as more

emphasized than repeated verbs among the bridging items.

The experimental results constitute an empirical challenge for the prior deaccenting literature. Proponents of both the grammatical licensing mechanism and the accommodation mechanism mark the canonical examples of deaccenting under inference as categorically acceptable, and a major goal of both classes of model is to generate felicitous deaccenting under both identity and inference. In both cases, deaccenting under inference is meant to be a straightforward operation, so it is surprising that there was no evidence that inferable constituents were deaccented in the analysis of the aggregate data for Experiments 9 and 10.

However, as discussed above, the lack of deaccenting of inferable constituents appears to be more problematic for the grammatical licensing accounts than the accommodation ones. In the grammatical accounts, the licensing of deaccenting of identical and inferable constituents are given the same treatment. Antecedent-identical constituents are treated merely as a subset of inferable constituents, as the existential closure of any constituent entails itself. In addition, these accounts generally mandate deaccenting of any constituent that can be deaccented through constraints such as *AVOIDF* or *MAXIMIZE BACKGROUND*. Reconciling a requirement for deaccenting with the lack of such behavior on inferable constituents in the production study would appear to require that the speakers were unaware of the inferencing relations linking the antecedent verbs and the target inference verbs. This seems unlikely in light of the high inferability scores in the inference condition from Experiment 8, the norming study, but the possibility is discussed more below.

By contrast, the accommodation licensing mechanism accounts at least leave room for differential licensing behavior between identical and inferable constituents. In these accounts, the core grammatical relation licensing deaccenting is identity, which aligns well with the consistent deaccenting of repeated constituents in the production study. Deaccenting of inferable constituents is licensed through the accommodation of a covert antecedent containing constituents identical to any deaccented target constituents. This suggests that deaccenting of inferable constituents may be at least slightly costly compared to deaccenting of identical

constituents. It also leaves open the possibility that deaccenting inferable constituents is optional, while deaccenting identical constituents is likely mandatory.

However, it is difficult to reconcile a notion of optional deaccenting with the discursive function of accent; prosodic emphasis is tightly linked to discourse-novelty, and de-emphasis to discourse-accessibility, so it is not immediately apparent why one type of constituent should optionally be treated as either discourse-new or discourse-old. One possibility is that participants in the production study, not having planned the utterances themselves, assumed that the inferable verbs were actually meant to contrast with their antecedents. For instance, given an antecedent of *hugged*, participants may have concluded that the new choice of lexical item was meant to signal some contrast in meaning; if the second verb was meant to convey the same meaning as the antecedent, it would be more rational to use the same word. Under this interpretation, it is not surprising that the inferable verbs were accented, as they should carry contrastive focus.

This possibility recalls the major caveat regarding the production study that was raised in the introduction to this chapter. Since the speakers did not plan their own utterances, it is difficult to know how much to conclude about naturally occurring speech from the results of the production study. One possibility is the one mentioned immediately above, where participants attributed discourse-novelty to the inference verbs and produced them with contrastive focus, whereas a real-world speaker might deaccent an inferable verb in the same situation if they planned the utterance themselves and meant to highlight the givenness of the second verb.

It is also possible that the Experiment 9 participants performed the production task in a shallow way, reading the sentences aloud without accessing all the inference relations linking the different constituents. If participants did not realize that the inference verbs were inferable from their antecedents, it is not surprising that they would be realized similarly to discourse-new verbs, whereas a real-world speaker who planned their own utterance would likely be aware of inference relations linking the antecedent and deaccenting target.

Thus, the results of Experiments 9 and 10 are far from conclusive evidence against the deaccenting of inferable constituents. There are many potential explanations for why the Experiment 9 speakers might not have deaccented the entailment and bridging verbs that maintain a mandate for deaccenting of inferable constituents in naturally occurring speech. It is safest to conclude that these experiments merely failed to detect evidence of such deaccenting without deciding whether this is due to experimental limitations or any actual rarity of such deaccenting in naturalistic production.

Two types of data are likely to be more informative on the question of how readily the deaccenting of inferable constituents is licensed. The first is naturally occurring speech data, which is not subject to the limitations of Experiment 9 outlined above. The second is perception data, where listeners judge the felicitousness of canonical productions of accented and deaccented constituents that appear in an environment where they are inferable from an antecedent. This perception data avoids the problems of speakers having a choice not to deaccent or being unaware of inferencing relations, as it forces listeners to provide a rating of sentences given that the speaker has already chosen to deaccent an inferable constituent. As it is not immediately practical to gather a large amount of naturally occurring productions of inferable constituents, although a large-scale corpus project with an appropriate annotation system is conceptually feasible. Thus, the former option is considered to be outside the scope of this dissertation, whereas the latter type of experiment is the focus of the next chapter.

CHAPTER 7

PERCEPTION OF DEACCENTING UNDER IDENTITY AND INFERENCE

The previous chapter investigated the prosodic encoding of discourse-new, inferable, and repeated verbs. The results indicated that speakers did not deaccent inferable verbs. Inferable and new verbs were pronounced with high phonetic correlates and judged as emphasized by listeners, while repeated verbs had low correlates and were judged as not emphasized. This finding was judged to be a problematic empirical generalization for all of the theories of deaccenting licensing reviewed the chapter before. This is because each of these theoretical accounts takes the canonical examples of deaccenting under nonidentity to be acceptable; that is, the accounts are designed to generate the data. Thus, it was considered to be surprising that the inferable verbs in the production study were not deaccented, since the inference relations linking them to their antecedents were quite strong, as judged in Experiment 8. The results were argued to be particularly problematic for the grammatical account of deaccenting under nonidentity, which treats repeated and inferable constituents identically in the grammar.

However, the discussion in the previous chapter did not take the observations from the production study as conclusive evidence against deaccenting of nonidentical constituents. This is because it was possible that the lack of deaccenting under nonidentity was an artifact of the production task itself. Since participants did not plan the recorded utterances themselves, they may have performed the production task at only a superficial level. In particular, if they did not realize that inference verbs were linked by an inferencing relation to their antecedent, then it is not surprising that these verbs were not deaccented. The chapter ended with the observation that a perception study would be more useful for determining whether the deaccenting of inferable constituents was readily licensed.

This chapter takes on the investigation of deaccenting using perception studies. Unlike

the production study, where participants were given minimal guidance and could perform the entire task without producing any deaccented inferable material, a perception study allows for guaranteed exposure to the critical phenomenon. Each experiment in this chapter uses manipulated stimuli constructed from recordings generated using the paradigm from Experiment 9. Second-clause new and repeated verbs in this paradigm should canonically be accented and deaccented, respectively. Here, recorded second clauses with new or repeated verbs were cross-spliced to follow all of the different first clauses, so verbs with canonical phonetic realization of accentuation or deaccentuation could be new, inferable, or repeated. This allows for participants to rate the naturalness of accented and deaccented inferable verbs, and for comparison of these ratings to those for accented and deaccented new and repeated verbs.

Experiment 11 elicits ratings of these sentences with no additional support, akin to encountering the sentences in an out-of-the-blue context. Deaccenting of inferable material is predicted to be felicitous by both accounts of licensing; even the accommodation account predicts that when listeners encounter deaccented string-new material, they should accept the deaccenting if the constituent is inferable from an antecedent (Fox, 2000). However, as in the previous chapter, inferable verbs in nuclear position actually patterned with new verbs, being judged as relatively infelicitous when deaccented and felicitous when accented. By contrast, repeated verbs sounded better when deaccented and worse when accented.

In light of this finding, the subsequent experiments focused on perception in contexts that supported an identical reading for inferable verbs and their antecedents. Experiments 12 and 13 asked participants to listen to the sentences after reading a context describing a situation supporting “situationally identical” readings for the two verbs. In these contexts, the previous preference of accenting (nuclear) inferable verbs, rather than deaccenting them, diminished in magnitude compared to when the sentences were heard out of the blue. Experiment 14 replicated Experiment 11, but added the presupposition trigger *too* to the end of every stimulus. Here, there was a preference for deaccenting inferable verbs, and it was

inferable and repeated verbs that patterned together to the exclusion of new verbs.

Together, the results from this chapter indicate that deaccenting of inferable constituents can be acceptable, but it is more restricted than suggested in the literature. In the cited examples, it is often suggested that a constituent merely being lexically inferable from an antecedent is sufficient to license deaccenting. However, the experiments here show that this is not the case, at least for the sentence types and constituents currently under investigation. Rather, deaccenting of antecedent-nonidentical constituents seems to require additional contextual support beyond the antecedent and target themselves, such as from the broader discourse context or a presupposition trigger. The acceptability of deaccenting appears to increase as a function of this contextual support.

The general discussion below argues that the experimental results are more compatible with the accommodation account of licensing for deaccenting of inferable constituents than with the grammatical account. The grammatical account does not leave a clear path for felicitousness differences between deaccented repeated and inferable constituents, when empirically deaccented repeated constituents were uniformly accepted and the acceptability of deaccented inferable constituents depended on the context. The grammatical account is left needing to suggest that participants were unaware of the lexical inferencing relations or that calculating these relations is somehow costly. By contrast, the accommodation account does leave room for such an acceptability gap, since it schematizes deaccenting antecedent-nonidentical constituents as ungrammatical. The distribution of acceptable deaccenting is more restricted than presented in the accommodation literature, but there is a clear path forward for this account. Although the accommodation required to license deaccenting is evidently not available in out-of-the-blue contexts, the contextual manipulations used in the experiments appear to facilitate the operation.

7.1 Experiment 11: Perception of emphasis in out-of-the-blue contexts

The goal of Experiment 11 is to conduct a controlled investigation of native listeners' judgments of the felicitousness of both accentuation and deaccentuation on constituents with various discourse statuses, and in particular on inferable constituents. In order to develop a full factorial design for the experiment, manipulated tokens from the production study from the previous chapter, as well as replications of it, were used. While manipulating stimuli for a perception study is less than ideal, it is not clear that speakers, even if prosodically trained, would be able to pronounce accented or deaccented constituents in a naturalistic way in an unusual discourse context (e.g., to deaccent a verb when it was actually new).

The experiment exposes participants to cross-spliced recordings of sentences in the style of those elicited in Experiment 9, of the form *SVO and SVO*. Certain of these production trials should elicit canonically accented or deaccented tokens of the second-clause verb – when it is discourse-new or repeated, respectively. By splicing together the first and second clauses of different production trials, then, recordings can be constructed with any combination of verb status and emphasis. For instance, a verb that was discourse-new in production, which should be accented, could be placed after a first clause that makes it repeated with respect to the cross-spliced string.

Thus, cross-splicing allows for participants to rate both accented and deaccented verbs with three different discourse statuses: discourse-new, inferable, and repeated. Comparing the results for the inferable conditions to those for the new and repeated conditions will lead to conclusions about the relative felicitousness of the deaccenting of inferable constituents – and in particular, its felicitousness relative to the deaccenting of antecedent-identical constituents. This comparison has the potential to disambiguate between the grammatical and accommodation licensing accounts for deaccenting, as they make different predictions about whether deaccenting of identical and inferable constituents must be equally licensed.

7.1.1 *Design and materials*

The critical manipulation in this experiment was to expose participants to recordings of verbs that were produced as discourse-new or repeated, which canonically are expected to be accented and deaccented, respectively, in contexts where they have different apparent discourse statuses. This was accomplished by cross-splicing recordings of new- and repeated-verb productions of the second clause from Experiment 9-like stimuli so that they followed recordings of first clauses making the verb either discourse-new, inferable, or repeated.

The experiment had a 3 (verb status) x 2 (object status) x 2 (accent status) manipulation. Verb status refers to the verb relation that exists in the actual (manipulated) sentence rated by the participants; the possible levels are new, inferable, and repeated, as in Experiment 9. Again, the inference relation relevant to the “inferable” condition is entailment in half the items and implicational bridging in the other half.

The possible levels for object status are new and old, as in the experiments in the previous chapter. As a reminder, new-object productions generally carry nuclear pitch accent on the second-clause object, which was seen in the previous chapter largely to neutralize phonetic differences between the realizations of preceding verbs with different discourse statuses. In old-object productions, there is the possibility for a nuclear pitch accent on the second-clause verb, which yielded greater phonetic contrast between accented and deaccented verbs.

For accent status, the possible levels are accented and deaccented, which refers to the canonical prosodic realization of the verb expected for the production from which the second-clause recording was taken. “Accented” means that the verb was discourse-new in its original production condition, while “deaccented” means that the verb was repeated in its original production condition.

While Experiment 9 featured 12 items (6 entailment items and 6 implicational bridging items), Experiment 11 used the full set of 48 verb triples that were selected after the completion of Experiment 8. As a result, it was necessary to generate recordings of the 36 new items in each of the production conditions for use in Experiment 11.

To record these stimuli, two speakers – one female, one male – who had participated in Experiment 9 returned to the lab for three additional recording sessions. These two speakers were judged after Experiment 9 to have been particularly expressive with their prosody on the basis of both impressionistic assessment of the recordings and investigation of their production of phonetic correlates of accent. They were selected in hopes that their productions of accented and deaccented verbs would be highly distinct. In each recording session, the participant repeated the same procedure from Experiment 9, but with a new set of 12 items swapped in to the experiment. The full set of stimuli recorded by the participants is shown in Appendix L.

After recording the full set of production stimuli, the recordings were cross-spliced to create the perception stimuli for the current experiment. The two speakers were generally reliable in producing the sentences with a substantial pause between the first SVO clause and the word *and*, so the recordings were split in this pause to avoid obvious artifacts of the manipulation. The recordings were clipped at zero crossings on both ends to avoid phonetic artifacts in the cross-spliced recordings.

For each object level for a particular item (new object and old object), the second clause of the new-verb production and the repeated-verb production were isolated as the “accented” and “deaccented” second-clause productions, respectively. These were then cross-spliced with the first clauses for the other productions with the same object status. For instance, the second clause of a new verb, new object production was cross-spliced with the first clause of a repeated verb, old object production. The goal of this operation is to create a recording where the apparent inferencing relation linking the two verbs is identity – since the pronounced verb is identical in both clauses – but the prosody of the second clause is consistent with the second verb being new (i.e., accented), since the second clause recording was taken from a new-verb production.

Note that in some cells of the design, cross-splicing was not necessary. For instance, in the condition where the second-clause verb should be discourse-new in production, so it

would be accented, and should be discourse-new in its surface form, the original production recording was used rather than a recording made by cross-splicing two halves of the same recording.

By cross-splicing both new- and repeated-verb productions after the different possible first clauses in this way, 6 different conditions were created for each of the new-object and old-object halves of an item. Critically, the second-clause verb could be canonically accented (if it was new in production) or deaccented (if it was repeated in production), but in its surface position, it could be discourse-new, inferable, or repeated. This allows for the assessment of the felicitousness of both types of prosody in contexts where they might not naturally be produced, e.g., a deaccented discourse-new verb.

Table 7.1 shows a set of sample stimuli for one experimental item. The curved arrows in this table show the sources and destinations of cross-spliced second clauses. For instance, on the second line (accented/inference/new), the first clause recording is from the original inference verb/new object production trial, while the second clause that was spliced after it is from the original new verb/new object trial. Recall that stimuli were constructed for each item for both a male and female speaker. Recall also that from Experiment 8, the mean inferencing scores out of 7 for the verb pairs used in the full set of stimuli are: entailment/inferable: 6.49; entailment/discourse-new: 2.12; bridging/inferable: 5.78; bridging/discourse-new: 2.15.

7.1.2 Procedure

The experiment was conducted using Amazon Mechanical Turk and Ibex Farm. Participants were recruited on AMT with a HIT noting that they were required to be native speakers of English over the age of 18 and using a computer in the United States. When they accepted the HIT, they followed a link to the experiment hosted on Ibex Farm, provided informed consent, completed a demographic survey, read the instructions for the experiment, and then completed the experiment. At the end of the experiment, participants received a code to enter on AMT to verify completion of the experiment. Payment of 1.00 USD was made

Accent status	Verb status	Object status	First clause	Second clause
Accented	New		Sebastian punished Justin,	and Nan surprised Amy.
	Inference	New	Bradley Astounded Justin,	and Nan surprised Amy.
	Repeated		Julian surprised Justin,	and Nan surprised Amy.
Accented	New		Gabriel punished Amy,	and Nan surprised Amy.
	Inference	Old	Ethan astounded Amy,	and Nan surprised Amy.
	Repeated		Benjamin surprised Amy,	and Nan surprised Amy.
Deaccented	New		Sebastian punished Justin,	and Nan surprised Amy.
	Inference	New	Bradley Astounded Justin,	and Nan surprised Amy.
	Repeated		Julian surprised Justin,	and Nan surprised Amy.
Deaccented	New		Gabriel punished Amy,	and Nan surprised Amy.
	Inference	Old	Ethan astounded Amy,	and Nan surprised Amy.
	Repeated		Benjamin surprised Amy,	and Nan surprised Amy.

Table 7.1: Sample stimuli for Experiment 11. Curved arrows indicate source and destination for cross-spliced second-clause recordings.

through the AMT system.

The instructions for the experiment familiarized the participants with the experimental task and were followed by two practice trials using the experimental setup. Participants also verified that sound was playing at an appropriate volume during the practice trials.

On each trial, participants saw a “warning screen” that displayed as text the sentence they were about to hear. Upon pressing a key, the warning screen changed to a response screen and the recorded sentence automatically played. After listening to the full sentence, participants used the number keys on their keyboard or clicked on a response on the screen to indicate how natural they found the production of the sentence, where 7 was the most natural and 1 was the least natural. The instructions told participants to go with their first instinct if they were unsure of their response. They also instructed participants to focus on the “melody or tune” of the sentence. This phrase was meant to encourage participants,

The sentence you will hear is:

Sebastian punished Justin, and Nan surprised Amy.

Press any key to hear the sentence.

On a scale from 1 to 7, where 1 is the least natural and 7 is the most natural, how natural did you find the "melody" or "tune" of this sentence?

Less natural
 1
2
3
4
5
6
7
More natural

Figure 7.1: Sample screens from Experiment 11. Top: Warning screen. Bottom: Response screen, during which recorded sentence *Sebastian punished Justin, and Nan surprised Amy* plays.

informally, to focus on the prosody of the sentence rather than other features (such as whether the speaker had an accent).

The experimental trials were organized into 48 lists of 24 critical trials. The lists were constructed so that each participant was exposed to each item and each of the levels for verb status, object status, accent status, and speaker gender an equal number of times, and no participant saw the same combination of these multiple times. Participants were sequentially assigned to lists in the order in which they began the Ibex Farm portion of the experiment.

In addition to the critical experimental items, each list contained the same 10 filler items. These filler items were intended to have clear naturalness responses given the experimental task (i.e., have clearly felicitous or clearly infelicitous prosody) so they could be used to remove the data from inattentive participants from the analysis. The order of presentation for critical and filler trials was randomized for each participant.

Figure 7.1 shows the layout of a sample trial from Experiment 11, with the warning screen at the top and the response screen (during which the recorded sentence played) at the bottom.

7.1.3 *Participants*

142 participants took part in the experiment. 67 of the participants were female, and 75 were male. The mean age of the participants was 36.6 years. One participant’s data was excluded from the analysis because they failed to identify themselves as a native speaker of English in the demographic survey. A further 5 participants were excluded from the analysis because their mean response time was under 1000 ms, indicating that they were not fully attending to the experiment. Among the participants whose data was analyzed, the mean completion time was 8 minutes and 1 second. Participants received USD 1.00 as compensation for completing the experiment.

7.1.4 *Predictions*

The grammatical and accommodation accounts of emphasis licensing make similar predictions for the results of this experiment, as they did for Experiment 9. Proponents of both models consider deaccenting under nonidentity straightforward, and as such, the models are designed to generate sentences with deaccented inferable material without much difficulty.

The grammatical models predict that (surface) discourse-new verbs should receive high ratings when accented, and relatively lower ratings when deaccented. Deaccenting of new constituents is not predicted by any of these accounts, as a constituent must be inferable from an antecedent constituent (modulo existential type shifting) to be deaccented grammatically. Conversely, these accounts predict that repeated verbs should receive high ratings when deaccented and low ratings when accented. Identical constituents are trivially entailed by their antecedent correlates, meaning that deaccenting is licensed under the unified accounts. Further, deaccenting is generally made mandatory when it is licensed by constraints such as AVOIDF (Schwarzschild, 1999), suggesting that it should not be felicitous to accent such constituents.

Critically, the grammatical accounts also predict that inferable constituents should be rated highly when deaccented and less highly when accented. Again, constituents can be

deaccented when their existential closures are entailed by the existential closure of a constituent in the antecedent. By design, the inferable verbs are made highly available by their antecedent verbs, as normed in Experiment 8. Thus, since the inferable verbs meet the grammatical requirement for deaccenting, they should sound natural when deaccented and relatively unnatural when accented.

Similar to the grammatical accounts, the accommodation approach to deaccenting licensing predicts high ratings for accented new verbs, low ratings for deaccented new verbs, high ratings for deaccented repeated verbs, and low ratings for accented repeated verbs. According to these accounts, the grammatical relation licensing deaccenting is one of string identity between the antecedent and the target constituent. As such, deaccenting should be licensed on repeated verbs, since they have an identical correlate in the antecedent, but not for new verbs, as they have no antecedent correlate (nor can a covert antecedent containing one be accommodated).

The accommodation approach also predicts high ratings for deaccented inferable verbs, and possibly lower ratings for accented inferable verbs. According to (Fox, 2000), a deaccented inferable verb constitutes *accommodation-seeking material*, since its deaccented status signals that it is discourse-old, but there is no correlate in the antecedent. Listeners, assuming the speaker communicating cooperatively and produces grammatical utterances, will seek to “repair” the utterance by accommodating a covert antecedent. This antecedent would contain an identical correlate for the deaccented target verb such that the deaccenting is grammatically licensed under identity with the covert antecedent. This operation should be possible for the inferable verbs, since it is by design reasonable to conclude that the inferable verb is the case when the antecedent is the case.

It is less clear that the accommodation approach predicts that accented inferable verbs should receive low ratings. Since the listener in such a trial is being presented with an accented verb, there is no accommodation-seeking material. Hearers of an accented inferable verb might conclude that it is accented merely because the grammatical requirement for

deaccenting is not fulfilled, or alternately because there is some shade of difference in the meaning of the antecedent and second-clause verbs such that the (otherwise-)inferable verb is felicitously treated as discourse new.

A critical difference between the grammatical and accommodation approaches that was considered in the last chapter is the possibility for differential behavior between (deaccented) new verbs and inference verbs. Because the grammatical approach uses a single mechanism to account for the deaccenting of both identical and inferable material, it should not be possible for the deaccenting of one type of material to be felicitous and the deaccenting of the other to be infelicitous. By contrast, the accommodation approach proposes that identity is the only grammatical driver of deaccenting, whereas the deaccenting of inferable material is due to a separate operation. This opens the door for the deaccenting of inferable material to be relatively less felicitous than that of identical material, since such deaccenting is, strictly speaking, ungrammatical. (However, note that the observation of infelicitous deaccenting of inferable material would run counter to the spirit of the accommodation approach, which also treats the deaccenting of inferable material as straightforward.)

Finally, it is worth recalling from the production study that there was virtually no difference in the realizations of discourse-new and repeated verbs in prenuclear position, i.e., when the second-clause object was discourse-new and carried a nuclear pitch accent. As such, it is reasonable to expect that listeners in Experiment 11 will be unable to distinguish the prosody of the different verb status conditions when the object is new. By contrast, there were clear phonetic and perceptual differences between new and repeated verbs when the verb was in a position to receive a nuclear pitch accent, i.e., when it was followed by a discourse-old, deaccented object. Thus, the theoretical predictions discussed above should be considered to hold only for the old-object conditions, where the “accented” and “deaccented” second-clause verb can reasonably be expected to be distinguishable. In all likelihood, there will be no differences by condition for the new-object conditions, since the “accented” and “deaccented” verbs are expected to be indistinguishable from one another.

7.1.5 *Results*

The results of Experiment 11 are shown in Figures 7.2-7.5. These figures highlight two different comparisons that are made in the analysis and discussion below.

Figures 7.2 and 7.3 directly compare sentences with the same accent status and object status, but different verb statuses, for the entailment and implicational bridging items, respectively. This highlights the effect of verb status on how natural a sentence with a particular verb emphasis – accented or deaccented – sounds to listeners.

Figures 7.4 and 7.5 compare sentences with the same verb status and object status and different accent statuses. This highlights the preference for either an accented or deaccented verb in sentences with the same discourse status throughout the second clause on both the verb and object.

A visual inspection of the old-object results suggests that participants did not treat deaccented inferable constituents as acceptable on par with deaccented repeated constituents. The verb status comparison suggests that new and inference verbs sound best when accented, while accented repeated verbs are not acceptable. Repeated verbs sound best when deaccented, while deaccented new and inference verbs were less acceptable. Interestingly, deaccented new and inference verbs received middling ratings; they do not appear to be as unacceptable as accented repeated verbs. Also, in the implicational bridging items, the difference between deaccented new/inference and repeated verbs is less apparent, probably because the ratings for repeated verbs were lower than for the entailment items.

In the accent status comparison, the plots suggest a reliable preference for accenting both new and inference verbs and a reliable preference for deaccenting repeated verbs. As expected, the plotted new-object results suggest that naturalness ratings were largely unaffected by both verb status and choice of accent.

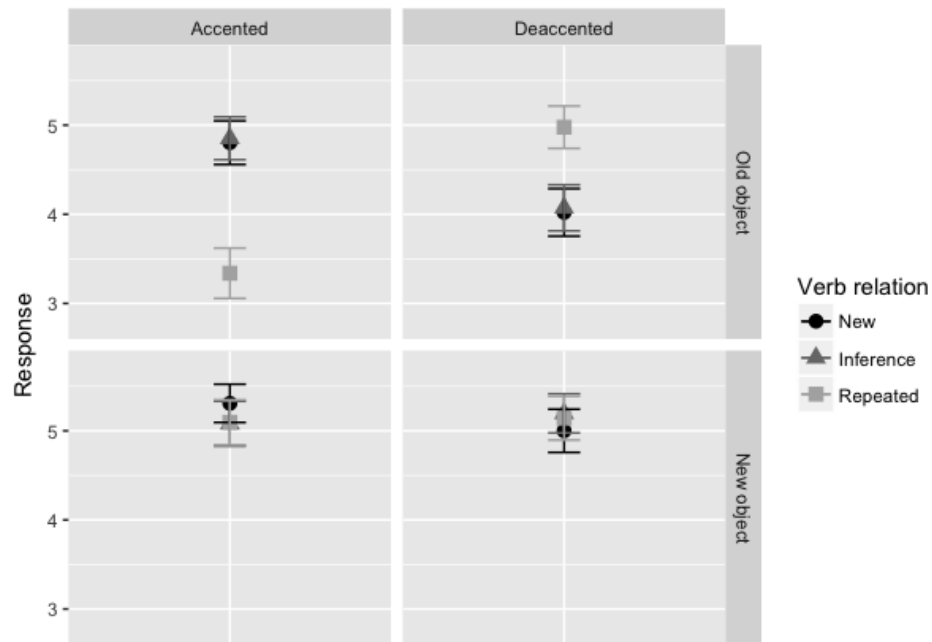


Figure 7.2: Experiment 11 results by verb status, entailment items. Error bars: 95% confidence interval.

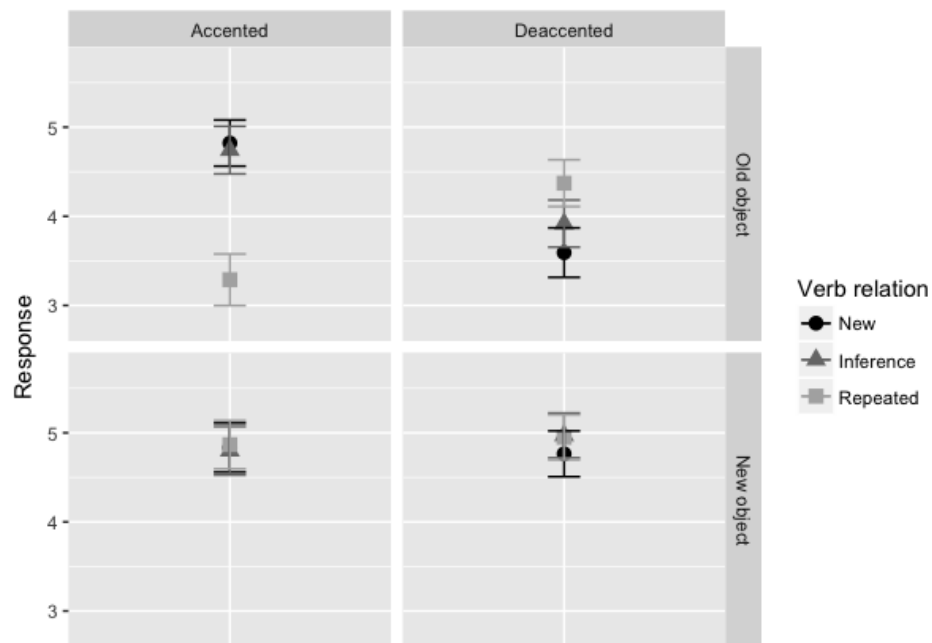


Figure 7.3: Experiment 11 results by verb status, implicational bridging items. Error bars: 95% confidence interval.

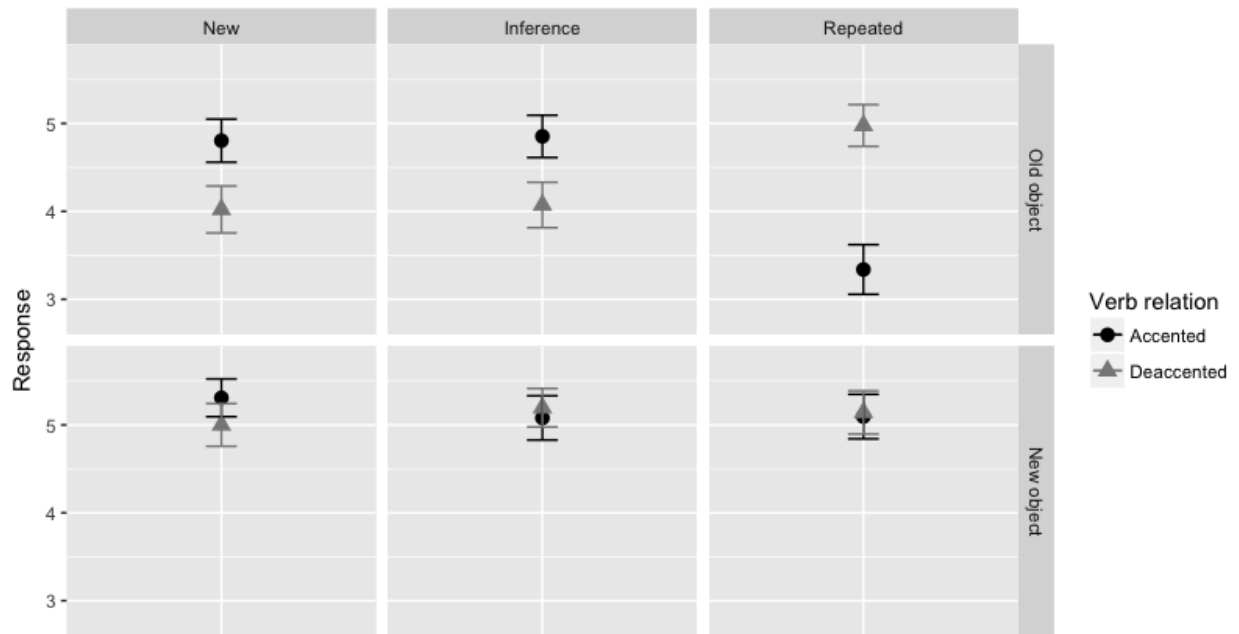


Figure 7.4: Experiment 11 results by accent status, entailment items. Error bars: 95% confidence interval.

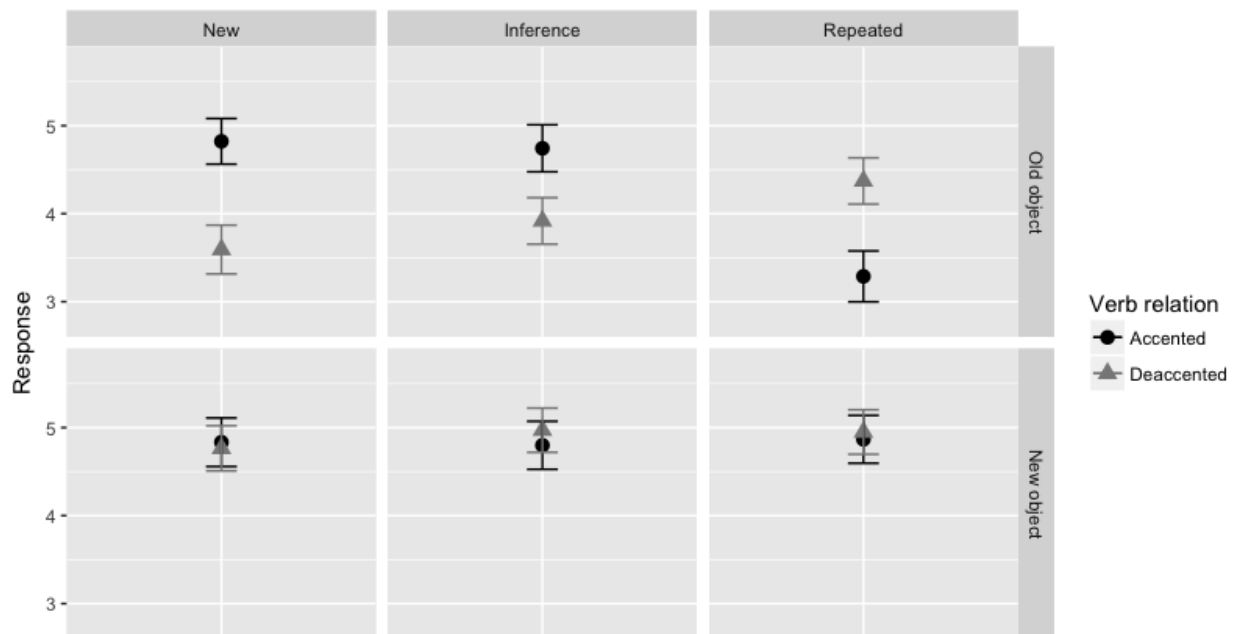


Figure 7.5: Experiment 11 results by accent status, implicational bridging items. Error bars: 95% confidence interval.

7.1.6 Analysis

Separate analyses were carried out for the entailment items and the implicational bridging items. For each set of items, a linear mixed-effects regression model was constructed with a three-way interaction of verb status, object status, and accent, plus each of the component two-way interactions and main effects and random effects for participant and item.¹ For each model, the three-way interaction was significant (p 's < .001).

On the basis of the significant three-way interactions, further analyses were carried out on subsets of the data. First, the data were separated by accent status to explore the effect of verb status on listener's assessment of the naturalness of an accented or deaccented verb token.

For each item set, separate linear mixed-effects models were constructed for the accented- and deaccented-verb trials. Each model had an interaction of verb status and object status, the component main effects, and random intercepts for participant and item.² For the accented/entailment, deaccented/entailment, and accented/bridging models, the effect of verb status, the effect of object status, and interaction were all significant (p 's < .001). For the deaccented/bridging models, the three effects were also significant, but less robustly so (verb p < .01, object p < .001, interaction p < .05).

Based on the significant verb-object interactions, paired comparisons were carried out within levels for object status and between levels for verb status. Table 7.2 summarizes the results of the paired comparisons for each model.

Additionally, the data were separated by verb status to explore the influence of the choice of verb emphasis on listener's perception of the naturalness of a token with a particular discourse status. For each of these subsets, a linear mixed-effects regression model was constructed with an interaction of accent status and object status, the component main

1. **lmer** model specification: $\text{Response} \sim \text{Verb} * \text{Object} * \text{Accent} + (1 \mid \text{Participant}) + (1 \mid \text{Item})$.

2. **lmer** model specification: $\text{Response} \sim \text{Verb} * \text{Object} + (1 \mid \text{Participant}) + (1 \mid \text{Item})$.

Item group	Accent status	Object status	New-Inference	New-Repeated	Inference-Repeated
Entailment	Accented	New	n.s.	n.s.	n.s.
Entailment	Accented	Old	n.s.	***	***
Entailment	Deaccented	New	n.s.	n.s.	n.s.
Entailment	Deaccented	Old	n.s.	***	***
Bridging	Accented	New	n.s.	n.s.	n.s.
Bridging	Accented	Old	n.s.	***	***
Bridging	Deaccented	New	n.s.	n.s.	n.s.
Bridging	Deaccented	Old	n.s.	***	*

Table 7.2: Experiment 11: Paired comparisons for verb status analysis. n.s. = not significant ($p > .1$). * = $p < .05$. *** = $p < .001$.

effects, and random effects for item and participant.³

The main effects of accent status and object status and their interaction were each significant for all six models. The p-values were all less than .001 with the following exceptions: entailment/new verb interaction, $p < .05$; entailment/inference verb accent status, $p < .01$; bridging/inference verb accent status, $p < .01$.

Based on the significant interactions, paired comparisons were carried out between the levels for accent status and within the levels of object status. Table 7.3 summarizes these paired comparisons.

7.1.7 Discussion

The analysis of the data separated by accent status yielded results that are similar to those for the production data from Experiments 9 and 10. There was always a significant interaction of verb status and object status, indicating that the effect of verb status on the naturalness ratings differed depending on whether the object was old or new.

As it turns out, when the object was new, no pair of verb statuses received significantly different ratings in either set of items, with either an accented or deaccented verb. Recall that in new-object productions, the second-clause object is expected to be realized with a

3. `lmer` model specification: $\text{Response} \sim \text{Accent} * \text{Object} + (1 \mid \text{Participant}) + (1 \mid \text{Item})$.

Item group	Verb status	Object status	Accented- Deaccented
Entailment	New	New	marg.
Entailment	New	Old	***
Entailment	Inference	New	n.s.
Entailment	Inference	Old	***
Entailment	Repeated	New	n.s.
Entailment	Repeated	Old	***
Bridging	New	New	n.s.
Bridging	New	Old	***
Bridging	Inference	New	n.s.
Bridging	Inference	Old	***
Bridging	Repeated	New	n.s.
Bridging	Repeated	Old	***

Table 7.3: Experiment 11: Paired comparisons for accent status analysis. n.s. = not significant ($p > .1$). marg. = marginal ($p < .1$). *** = $p < .001$.

nuclear pitch accent. In the analyses in the previous chapter, this was seen to neutralize potential prosodic contrasts between preceding verbs of different discourse status. Thus, it is not surprising that in perception, the apparent discourse status of the verb does not affect naturalness ratings; as “typical” productions of new, inferable, and repeated verbs are quite similar in these sentences, participants likely did not detect any anomalous prosody when a new second clause was spliced in. In addition, it is not clear that even if there were reliable phonetic differences on these verbs, participants would have strong intuitions about what prosodic realizations are acceptable in prenuclear position.

By contrast, the old-object conditions, where an accent on the verb can be nuclear, revealed an effect of verb status that is familiar from the previous chapter. In all old-object paired comparisons by verb status, the ratings for new versus repeated and inference versus repeated verbs were always significantly different, while the ratings for new and inference verbs were never significantly different. Examination of the mean scores in Figures 7.2 and 7.3 suggests that sentences with an accented verb received significantly higher ratings when the verb was discourse-new or inferable compared to when it was repeated. The opposite holds for deaccented-verb sentences, which received significantly higher ratings when the

verb was deaccented than when it was new or inferable.

The data separated by verb status shows a similar pattern. Again, there was always a significant interaction of accent and object status, suggesting that the preference for an accented versus deaccented verb depended on whether the object was old or new. Paired comparisons showed that there was never a reliable preference for either accent status when the object was new, which is consistent with the observation from the production study that verbs in pre-nuclear position have highly similar realizations that are not readily distinguished by listeners.

In the context of an old object, however, there were consistent preferences for one production accent pattern or the other. An accented verb (that is, one that was pronounced as discourse-new in its original production condition) was significantly preferred for new and inference verbs for both the entailment and implicational bridging items. Conversely, a deaccented (repeated in production) verb was preferred when the verb was repeated in the perception stimulus.

Thus, as in the production study, the results of Experiment 11 show that in perception of accentuation in out-of-the-blue contexts, inferable verbs pattern with discourse-new verbs to the exclusion of repeated verbs. Unsurprisingly, participants preferred sentences where a discourse-new verb was accented to those where discourse-new verbs were deaccented, and preferred sentences with deaccented repeated verbs to those with accented repeated verbs. Most critically, sentences with inferable verbs did not reliably behave differently from new-verb sentences. Participants preferred that inferable verbs be accented rather than deaccented with roughly the same effect strength as for new verbs, and the ratings for both accented and deaccented inferable verbs were analogous to the ratings for the corresponding new verbs and reliably different from those for repeated verbs.

Recall that in the discussion of the production study, it was noted that a perception study using manipulated stimuli would be more apt for assessing whether the deaccenting of inferable constituents is readily licensed. Whereas the speakers in Experiment 9 may

have had the option to perform the production task only superficially, ignoring the critical inference relations, listeners in the perception study are forced to contend with the stimulus as it is presented to them. In particular, when listeners encounter a deaccented inference verb, even if it may appear on its surface to be unlicensed, they must assess what motivation the speaker may have had for considering the deaccenting licensed. (Recall that in Fox's (2000) formulation, deaccented material that does not have an identical correlate in the antecedent is *accommodation-seeking material*.)

However, the results of Experiment 11 suggest that the behavior of inference verbs in the production study was not spurious. When presented with a deaccented verb that stood in a strong inferencing relation to the antecedent verb, listeners rated the sentence similarly to the analogous case where the verb was entirely discourse-new, in which case deaccenting is unlicensed. By contrast, listeners rated inference verbs that were accented – a prosodic choice generally associated with discourse-novelty – as sounding natural.

As noted in the discussion for the previous chapter, the finding that inference verbs pattern with discourse-new verbs to the exclusion of repeated verbs is somewhat problematic for both classes of accounts of deaccenting licensing. This is because both groups of theoreticians cast deaccenting of inferable material as straightforward and intend for their theories to generate such examples. However, the sum of evidence collected so far suggests that this is not the case, at least for the verb constituents and antecedents analyzed in the current experimental paradigm. Not only were inferable verbs not deaccented in production, but they were also seen to behave exactly like discourse-new verbs in perception, with a strong preference that they be accented instead of deaccented.

Further consideration of the theories presented in the literature, however, suggests that the results should be viewed as more problematic for the grammatical accounts of emphasis licensing than the accommodation accounts. The grammatical accounts of emphasis licensing use a single grammatical definition to encompass the determination of those constituents that can be deaccented under identity with an antecedent constituent and those that are inferable

from an antecedent constituent. Since deaccenting is generally treated as mandatory when it is possible, it is difficult to reconcile the differential behavior of repeated and inferable verbs in the experiments so far with this class of theoretical account.

By contrast, the accommodation accounts privilege deaccenting under identity as a grammatical operation, whereas deaccenting under nonidentity is treated as a process of pragmatic inference, namely of accommodation of a covert antecedent containing identical correlates of the deaccented material. Unlike the grammatical accounts, this class of model does, in principle, allow for differential behavior between repeated and inferable constituents. It is not clear that deaccenting of inferable material is mandatory in production, since it involves an extragrammatical operation.

However, the mere possibility for differing behavior for repeated and inferable constituents under the accommodation approach does not necessarily mean that this account is supported by the experimental results. Like the grammatical accounts, this approach is designed to generate felicitous examples of deaccenting under inference. In addition, the Experiment 11 stimuli contained clear examples of Fox's (2000) accommodation-seeking material, namely deaccented verbs that were discourse-new or inferable. The accommodation approach suggests that in the inferable conditions, listeners should have responded to the ungrammatical deaccenting of the verb by attempting to reconcile it with the previous context through accommodation. Since the strong inferencing relation linking the verbs should make it reasonable to accommodate a new antecedent with the target verb in place of the original antecedent verb, accommodation should be possible in the inference conditions and lead to higher ratings, but this prediction was not borne out.

Evidently, there is more work to be carried out to determine why deaccenting under nonidentity was difficult in both production and perception in the current experimental paradigm, but is considered to be straightforward in the canonical examples cited in the literature on prosodic emphasis. One possibility is that the accommodation account is essentially correct, but that lexical inferencing relations alone are not sufficient to license

an accommodation operation even when the target clause contains accommodation-seeking material. The experiments below explore the possibility that ratings for deaccented inferable material can be improved when the broader context supports identical readings for the deaccenting target and its nonidentical antecedent.

7.2 Experiments 12 and 13: Perception of emphasis in supportive contexts

Both Experiment 9, the production study, and Experiment 11, the perception study with cross-spliced stimuli, showed evidence that in simple *SVO and SVO* sentences, verbs that stand in a strong inferencing relation to an antecedent verb are accented similarly to fully discourse-new verbs, to the exclusion of repeated verbs. This was taken as preliminary evidence in favor of an accommodation model of prosodic emphasis licensing, whereby the deaccenting of inferable constituents is licensed through a process of extragrammatical reasoning rather than directly by the grammatical constraint that licenses deaccenting under identity. This is because unlike the grammatical licensing account, the accommodation account provides the possibility for differential licensing of deaccenting for identical and inferable constituents; that is, the accommodation account allows for the deaccenting of identical constituents to be more felicitous than that of inferable constituents, while the unified account predicts that they should exhibit roughly identical behavior.

One key element of the accommodation model is its appeal to pragmatic reasoning as a pathway toward deaccenting under nonidentity. In Fox's (2000) formulation, deaccented material that does not have an identical correlate in the antecedent is *accommodation-seeking material*. Listeners who encounter such material should assume that their interlocutor did not mean to utter a sentence with an ungrammatical prosodic realization, and so should attempt to accommodate an antecedent that contains an identical constituent that would have otherwise licensed the deaccenting. In the case of inferable constituents, this should be

straightforward, since these constituents (or rather, their existential closures) are entailed by the antecedent, so the accommodation should always be consistent with the state of affairs. By contrast, it is not always felicitous to perform such accommodation when the deaccented constituent is not inferable, since the information contained in that constituent may not be entailed by the present discourse context.

While this account was designed to generate deaccenting under nonidentity rather straightforwardly, the previous experiment suggested that accommodation of covert antecedents in response to deaccenting of nonidentical material does not occur so readily (if, indeed, this is the mechanism that is in play). The goal of the current experiment, Experiment 12, is to explore whether the presence of a supportive discourse context can facilitate this accommodation operation.

Presumably, listeners may take a wide variety of information into account when deciding whether to accommodate a covert antecedent. The previous experiment suggested that lexical inferability alone is not sufficient to license this operation, but additional supportive information from the discourse at large may facilitate accommodation. Experiments 12 and 13 modifies the design of Experiment 11 to include a context of utterance for each sentence that is designed to support the conclusion that the inferable verbs and their antecedents constitute comparable events in context – that is, that the two verbs are “pragmatically” or “situationally” identical.

The main question is whether participants find the deaccenting of inference verbs to sound more felicitous given a supportive context. If the context suggests that the actions described by the antecedent and target verbs are comparable, participants may be more likely to accommodate a covert antecedent with a correlate for the target verb, since the context supports marking the target verb as discourse-given. Improvement of the ratings for deaccented inferable verbs in this experiment would support the accommodation licensing account for deaccenting, as this model predicts that listener’s willingness to accept deaccented inferable material should be sensitive to the information available in the broad discourse

Here is a scenario:

As they did every year, the teachers worried about how the students would interact with each other on the first day of high school.

Press any key to hear a speaker talk about the scenario.

On a scale from 1 to 7, where 1 is the least natural and 7 is the most natural, how natural did you find the "melody" or "tune" of this sentence?

Less natural
 1
2
3
4
5
6
7
More natural

Figure 7.6: Sample screens from Experiment 12. Top: “Warning” screen, giving participants the context of utterance. Bottom: “Response” screen, where the sound file (e.g., *Veronica bullied Jane, and Kendall intimidated Roy*) played automatically upon loading and participants provided a rating.

context. By contrast, the grammatical licensing models predict that there should be no change in the behavior of inferable constituents, since according to these models, whether an inferable constituent can be deaccented is strictly a function of its semantic relationship to the overt antecedent (although the general discussion of this chapter discusses some possibilities for reconciling the grammatical account with such an operation).

7.2.1 Design, materials, and procedure

The design, materials, and procedures for Experiments 12 and 13 were identical to those for Experiment 11 with the exceptions noted here.

In Experiment 12, the warning screen from Experiment 11 was modified so that instead of giving participants a text preview of the sentence they would hear, it provided a one-sentence context in which the participant was to imagine the recorded utterance occurred. Figure 7.6 shows an example of the modified warning screen and the response screen used in Experiment 12.

After Experiment 12 was conducted, there was some concern that the results were not comparable to those of Experiment 11 due to differences in the paradigm. In particular,

Imagine you just heard someone say:

Like every year, there were a lot of clever maneuvers during the legislative session.

Next, press any key to hear the same person say:

Henry supported Emma, and Lee defended Daniel.

On a scale from 1 to 7, where 1 is the least natural and 7 is the most natural, how natural did you find the "melody" or "tune" of this sentence?

Less natural
 1
2
3
4
5
6
7
More natural

Figure 7.7: Sample screens from Experiment 13. Top: “Warning” screen, giving participants the context of utterance and a preview of the target sentence. Bottom: “Response” screen, where the sound file (e.g., *Henry supported Emma, and Lee defended Daniel.*) played automatically upon loading and participants provided a rating.

since participants saw a preview of the target sentence in the former experiment but not the latter, it is possible that they responded differently when they were encountering the critical constituents without warning. For instance, when participants see a preview of the sentence, they might realize in advance that there is an inference relation linking the antecedent and target verbs, whereas this might be more difficult to calculate while processing the content of the sentence for the first time.

To account for this, Experiment 13 altered the Experiment 12 warning screen to include both the discourse context and a preview of the upcoming sentence. The Experiment 13 screen also indicated that the participants should imagine that the context sentence and target sentence were uttered by the same speaker. Sample screens from Experiment 13 are shown in Figure 7.7.

As it turns out, the results for Experiments 12 and 13 were comparable, so they are discussed together below.

For both experiments, the context sentences were static by item. The goal of the context

sentences was to facilitate participants' inference that the antecedent and second-clause verbs in the inferable conditions constituted events which, in context, could be considered as pragmatically identical for the purposes of accommodation along the lines of the dual-mechanism model. The goal was to provide informal support for this conclusion without overtly manipulating the discourse status of either of the verbs. That is, the context sentence never contained either the antecedent verb or the second-clause verb, as this would make them discourse-given. Instead, the goal was to construct contexts that were loosely relevant to the two verbs and might facilitate the conclusion that the two verbs denoted equivalent events in context.

For example, for the trial shown in Figure 7.6, the context sentence is *As they did every year, the teachers worried about how the students would interact with each other on the first day of high school*. In the inference condition, the antecedent and second-clause verb pair is *bullied* and *intimidated*. While the context sentence does not make either of these verbs explicitly given, it may help construe them as denoting equivalent events, since what is made at-issue in the context sentence is whether students will have positive or negative interactions with one another. In this way, the context sentence may encourage participants to accommodate an identical antecedent for *intimidated* more readily than they would given only the lexical inferencing relationship linking the two verbs (where the change in lexical item may signal novelty).

The other major change relative to Experiment 11 was the elimination of 6 items each in the entailment and implicational bridging item groups. Since the constraints on the form of the context sentence were somewhat restrictive, it was difficult to find a context sentence that suggested pragmatic identity for every pair of antecedent and inference verbs from the previous experiments. To account for this, context sentences were constructed for 18 entailment items and 18 bridging items, giving sufficient leeway to discard those items with difficult-to-unify verb pairs while still retaining the majority of the experimental items.

The experimental items were still organized into 48 lists with the same protocol as in

Experiment 11. However, in Experiments 12 and 13, participants completed 10 filler trials and 18 critical trials, compared to 10 fillers and 24 critical trials in Experiment 11. The full set of experimental stimuli, indicating the included items and their context sentences, can be viewed in Appendix M.

7.2.2 Participants

144 participants took part in Experiment 12. Of these, 53 were female and 91 were male. The mean age of the participants was 33.5 years. The data from 2 participants was excluded from analysis because they failed to identify themselves as native speakers of English in the demographic survey. Data from a further 7 participants was excluded from analysis because these participants' mean response time per trial was under 1000 ms, indicating that they were not fully attending to the experimental task. Among participants whose data was analyzed, the mean time to complete the Ibex Farm portion of the experiment was 8 minutes, 28 seconds. Participants received USD 1.00 as compensation for completing the experiment.

145 participants took part in Experiment 13, of whom 53 were female and 92 were male. The mean age was 35.6. The data from 1 participant who did not self-identify as a native English speaker and 10 participants whose mean response times were under 1000 ms were excluded. The mean completion time for analyzed participants was 9 minutes, 20 seconds. The procedure for payment and payment amount was the same as for Experiment 12.

7.2.3 Predictions

The accommodation account posits that deaccenting under nonidentity is licensed by an extragrammatical operation, namely accommodation of a covert antecedent with a constituent identical to the grammatically unlicensed deaccenting target. One interpretation of this account is that the availability of deaccenting could be modulated by the extent to which the accommodation operation is supported by the discourse context in which the target

sentence is situated. That is, if the context makes it more reasonable to conclude that a target-identical antecedent is acceptable to accommodate, then the deaccented target should be judged as more acceptable.

In these experiments, the critical observation is whether deaccented inference verbs become more acceptable in a supportive discourse context. In Experiment 11, deaccented repeated verbs (in nuclear position) were rated as highly natural to the exclusion of deaccented inference and new verbs, which were less natural and roughly comparable to each other. If the accommodation account is correct, the ratings for deaccented inference verbs in Experiments 12 and 13, which feature a discourse context supporting an identical reading of the target and antecedent verbs, may increase. This may result in comparable high ratings for deaccented repeated and inference verbs to the exclusion of new verbs, or ratings for deaccented inference verbs that are intermediate between the ratings for new and repeated verbs.

7.2.4 *Results*

The results of Experiment 12 are shown in Figures 7.8-7.11. As in the previous section, Figures 7.8 and 7.9 highlight the comparison of verbs by their discourse status, with the same accent status in the context of objects of the same status. Figures 7.10 and 7.11 compare accented and deaccented verbs in sentences with identical verb status and object status.

A visual inspection of the results suggests substantial effects of the addition of a context sentence compared to the results for Experiment 11. Most striking are the results for deaccented verbs in the context of an old object. In Experiment 11, deaccented repeated verbs in such contexts were rated significantly better than deaccented inference or repeated verbs. Here, the difference between inference and repeated verbs appears to have been neutralized. However, it is not clear from the plotted results that inference and repeated verbs were reliably rated higher than deaccented new verbs. Notably, the results in the deaccented/old

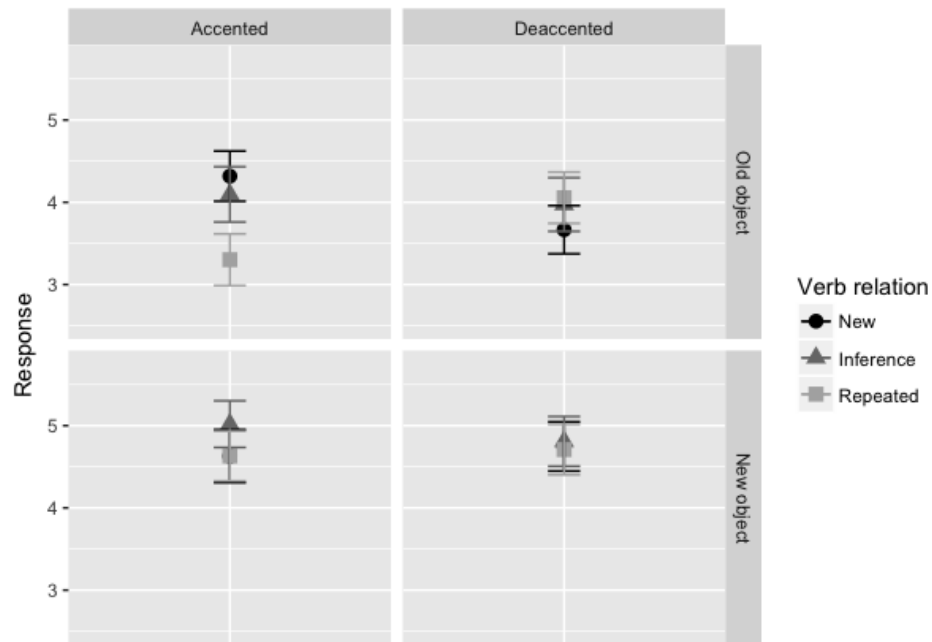


Figure 7.8: Experiment 12 results by verb status, entailment items. Error bars: 95% confidence interval.

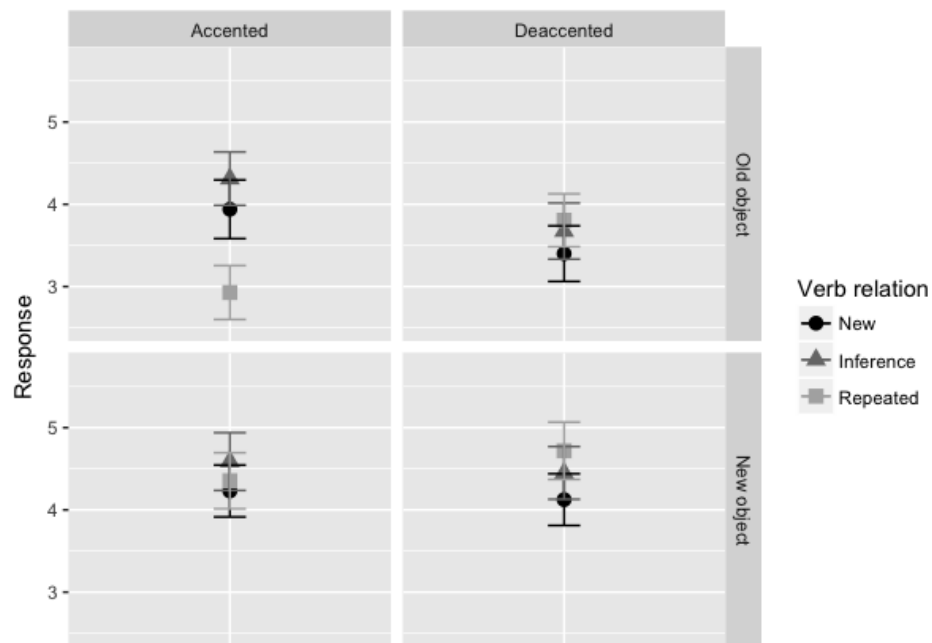


Figure 7.9: Experiment 12 results by verb status, implicational bridging items. Error bars: 95% confidence interval.

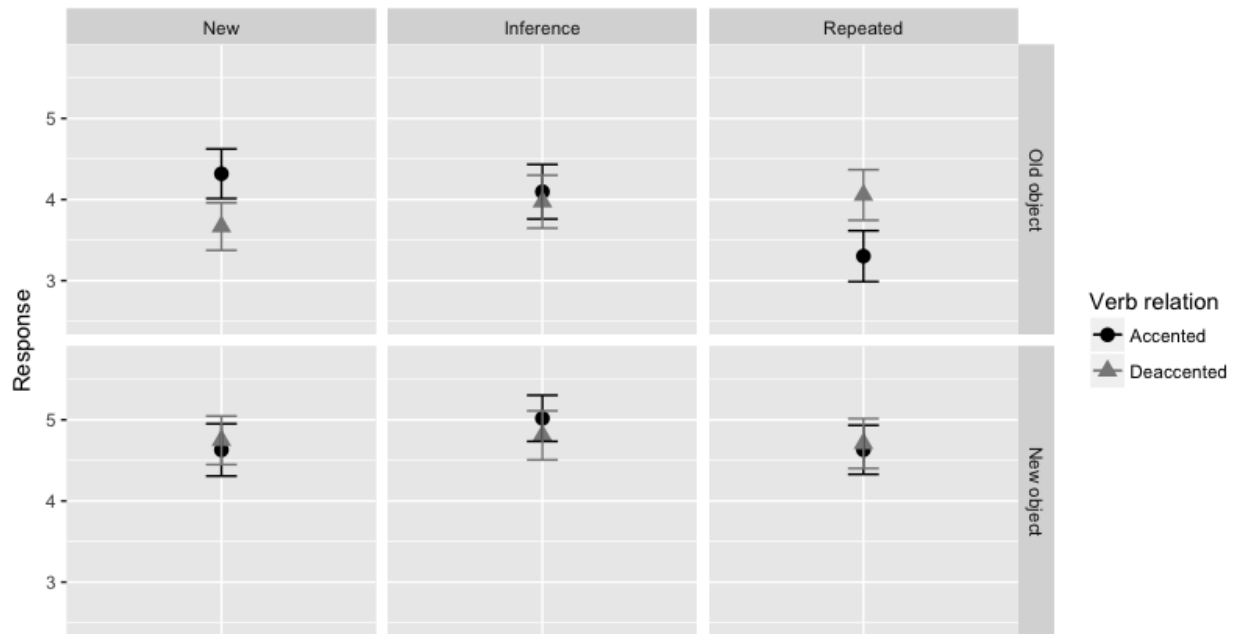


Figure 7.10: Experiment 12 results by accent status, entailment items. Error bars: 95% confidence interval.

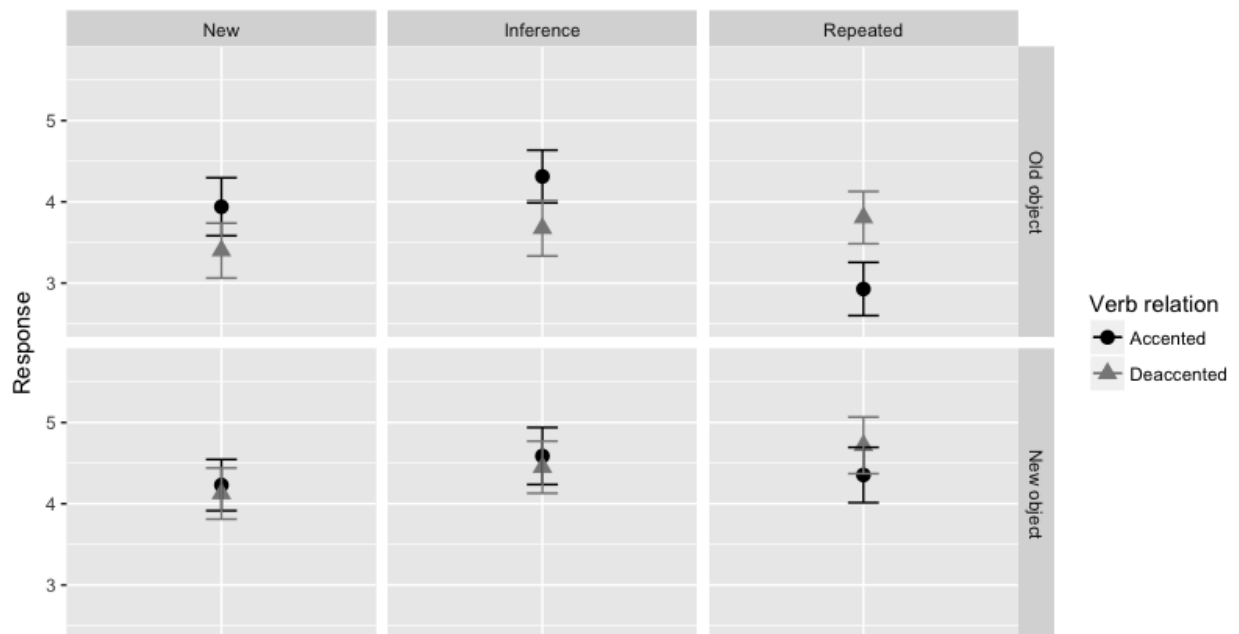


Figure 7.11: Experiment 12 results by accent status, implicational bridging items. Error bars: 95% confidence interval.

object conditions appear to have collapsed toward the middle of the rating scale relative to the results in the other conditions and from Experiment 11.

The results for accented verbs in the verb status comparison do not appear to have changed qualitatively relative to the results from Experiment 11. For accented verbs in the context of an old object, new and inference verbs appear to have been rated higher than repeated verbs, but not differently from each other. In the new-object conditions, there do not appear to be reliable differences in the ratings of verbs of different status, a common result for verbs in nuclear position.

Turning to the comparison of accented and deaccented verbs in identical contexts, new and repeated verbs appear to show the canonically expected results in nuclear position, with accentuation preferred for new verbs and deaccentuation preferred for old verbs. Interestingly, participants appear to have been indifferent between accentuation and deaccentuation for entailed verbs in nuclear position, whereas accented bridging verbs appear to be preferred over deaccented ones. There are no clear preferences between accentuation and deaccentuation in prenuclear position.

The results for Experiment 13 are shown in Figures 7.12 through 7.15.

The Experiment 13 results will not be discussed in detail, as they are largely similar to the results for Experiment 12. In general, the new-object conditions in Experiment 13 appear to have received higher scores among the bridging items than in the previous experiment, and there is less spread in the ratings for new-object sentences. Experiment 13 shows a clearer separation in the ratings for accented/new-verb and deaccented/new-verb sentences among old-object bridging items. Likewise, the accent distinction for old-object, inference-verb sentences is more pronounced for entailment items, but less pronounced for bridging items.

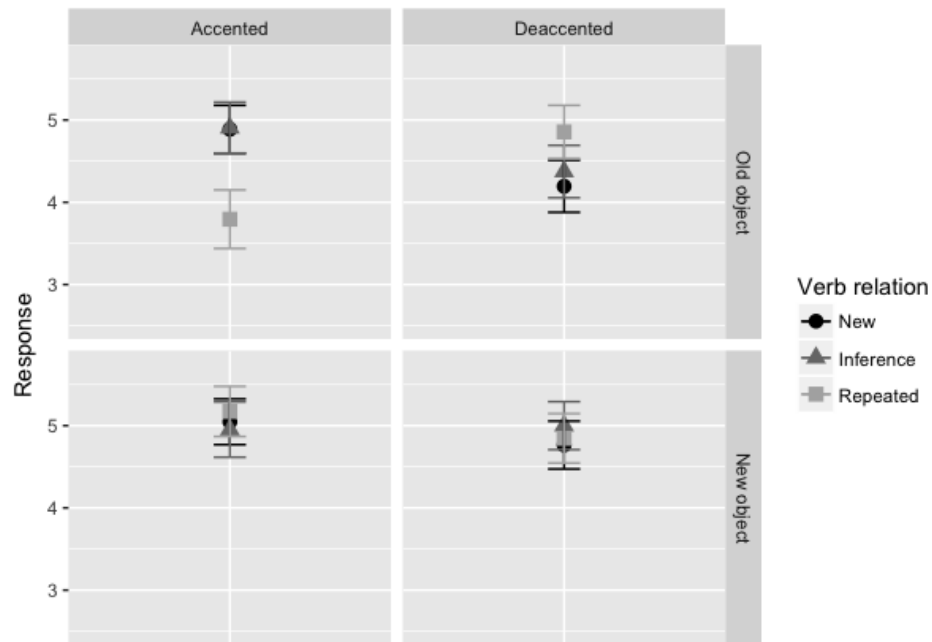


Figure 7.12: Experiment 13 results by verb status, entailment items. Error bars: 95% confidence interval.

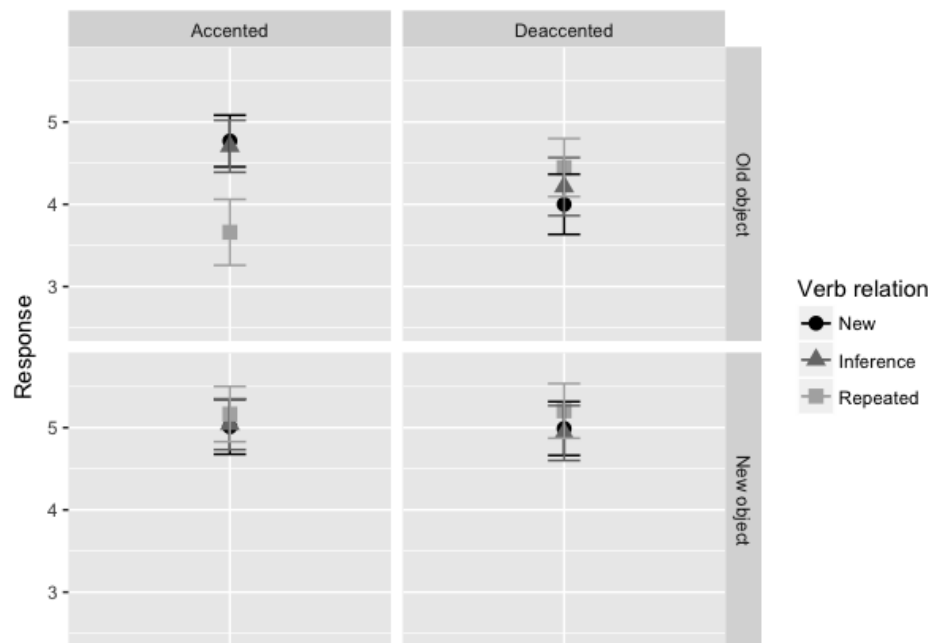


Figure 7.13: Experiment 13 results by verb status, implicational bridging items. Error bars: 95% confidence interval.

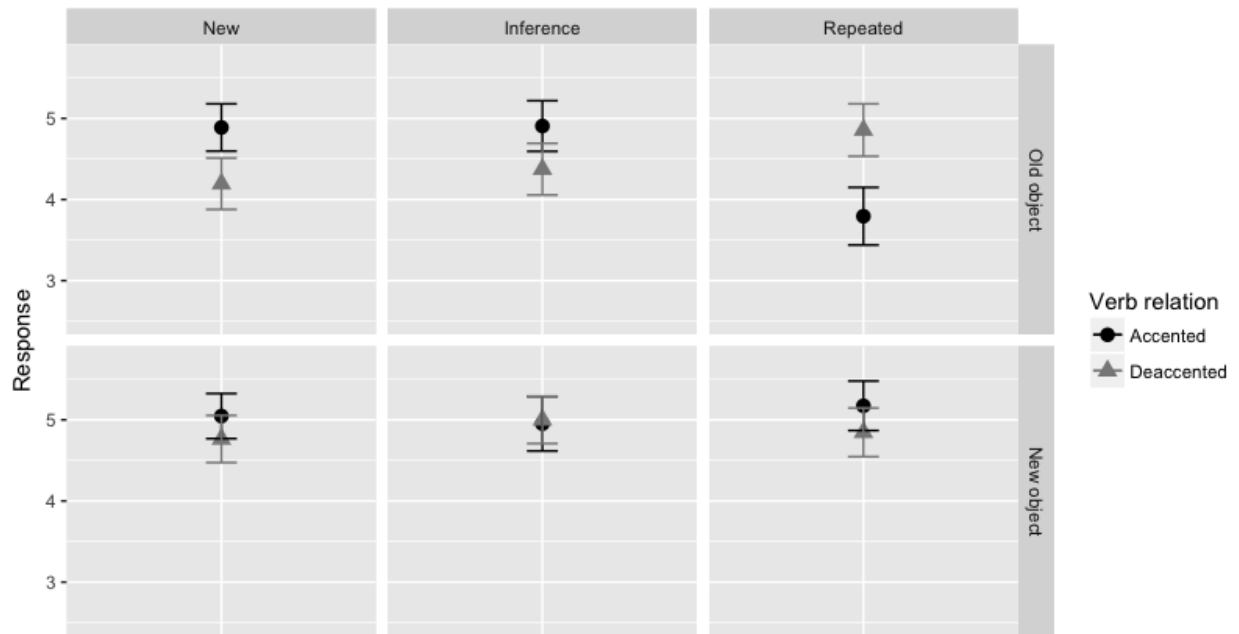


Figure 7.14: Experiment 13 results by accent status, entailment items. Error bars: 95% confidence interval.

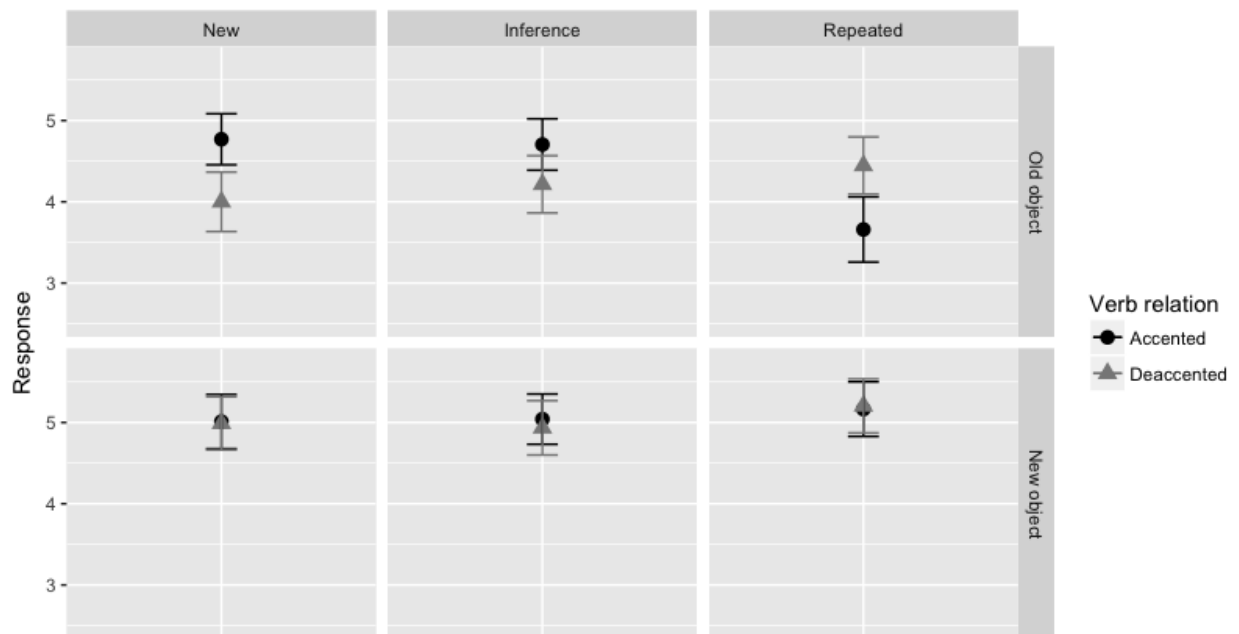


Figure 7.15: Experiment 13 results by accent status, implicational bridging items. Error bars: 95% confidence interval.

7.2.5 Analysis

The analytical approach to the results of Experiment 12 was identical to the one from Experiment 11. The three-way interaction of verb status, object status, and accent status was significant for both the entailment items ($p < .001$) and the bridging items ($p < .05$), so further analysis was conducted separately on subsets of the data.

The first models took subsets of the data by accent status, investigating the effect of verb status on the naturalness of either accented or deaccented verb sentences. Table 7.4 summarizes the results of this modeling.

Item group	Accent status	Verb status	Object status	Interaction
Entailment	Accented	***	***	**
Entailment	Deaccented	n.s.	***	n.s.
Bridging	Accented	***	***	***
Bridging	Deaccented	**	***	n.s.

Table 7.4: Experiment 12: Summary of models by inference relation and accent for verb status comparison. n.s. = not significant ($p > .1$). marg. = marginal ($p < .1$). ** = $p < .01$. *** = $p < .001$.

Where there was a significant verb-object interaction, paired comparisons were carried out by verb status within levels for object status. Where only the main effect of verb status was significant, but not the interaction, paired comparisons were carried out by verb status without reference to object status. When there was a main effect of object status but no significant interaction, paired comparisons were not carried out, since there were only two levels for object status. Table 7.5 summarizes the results of the paired comparisons for the verb status analysis of Experiment 12.

Next, the results were separated by the verb's discourse status (and item group), primarily to determine whether, for each status, whether accentuation or deaccentuation was rated as significantly more natural. This analysis corresponds to the visualization in Figures 7.10 and 7.11. Linear mixed models with main effects of accent status and object status and their interaction, and random intercepts for participant and item, were constructed for each

Item group	Accent status	Object status	New-Inference	New-Repeated	Inference-Repeated
Entailment	Accented	New	n.s.	n.s.	marg.
Entailment	Accented	Old	n.s.	***	***
Entailment	Deaccented	All data	No verb effect		
Bridging	Accented	New	n.s.	n.s.	n.s.
Bridging	Accented	Old	n.s.	***	***
Bridging	Deaccented	All data	n.s.	**	n.s.

Table 7.5: Experiment 12: Paired comparisons for verb status analysis. n.s. = not significant ($p > .1$). ** = $p < .01$. *** = $p < .001$.

Item group	Verb status	Accent	Object status	Interaction
Entailment	New	*	***	**
Entailment	Inference	n.s.	***	n.s.
Entailment	Repeated	**	***	*
Bridging	New	*	***	n.s.
Bridging	Inference	**	***	n.s.
Bridging	Repeated	***	***	marg.

Table 7.6: Experiment 12: Summary of models by inference relation and verb relation for accent status comparison. n.s. = not significant ($p > .1$). marg. = marginal ($p < .1$). * = $p < .05$. ** = $p < .01$. *** = $p < .001$.

subset of the data. The models are summarized in Table 7.6.

When there was a significant interaction between accent status and object status, paired comparisons were carried out to determine under what conditions there was a significant preference for accentuation or deaccentuation. When only main effects were significant, paired comparisons were not carried out, as both accent and object status have only two levels. Table 7.7 summarizes the paired comparisons.

The analytical approach for Experiment 13 was identical to that for Experiment 12. Tables 7.8 through 7.11 summarize the analysis for this experiment. In the initial models, the three-way interaction of verb status, object status, and accent was significant for both the entailment ($p < .001$) and the bridging items ($p < .01$).

Item group	Verb status	Object status	Accented-Deaccented
Entailment	New	New	n.s.
Entailment	New	Old	***
Entailment	Inference	All data	not tested (interaction n.s.)
Entailment	Repeated	New	n.s.
Entailment	Repeated	Old	***
Bridging	New	All data	not tested (interaction n.s.)
Bridging	Inference	All data	not tested (interaction n.s.)
Bridging	Repeated	New	marg.
Bridging	Repeated	Old	***

Table 7.7: Experiment 12: Paired comparisons for accent status analysis. n.s. = not significant ($p > .1$). marg. = marginal ($p < .1$). *** = $p < .001$.

Item group	Accent status	Verb status	Object status	Interaction
Entailment	Accented	***	***	***
Entailment	Deaccented	marg.	***	marg.
Bridging	Accented	**	***	***
Bridging	Deaccented	marg.	***	n.s.

Table 7.8: Experiment 13: Summary of models by inference relation and accent for verb status comparison. n.s. = not significant ($p > .1$). marg. = marginal ($p < .1$). *** = $p < .001$.

7.2.6 Discussion

The discussion will largely ignore the new-object results. As usual, there were few theoretically interesting findings in these conditions, with verb status having very little effect on the naturalness of sentences with various prosodic contours when the verb is in prenuclear position.

Experiment 12: For the old-object trials, where the verb was in nuclear position, the results for accented verbs were roughly comparable to the previous experiments in both production and perception. In both item groups, accented new and inference verbs both received higher ratings than accented repeated verbs, but their ratings were not significantly different from one another. This result is not particularly surprising, although it is worth pointing out that to the extent that the addition of a context sentence made inference verbs more “given,” this did not qualitatively diminish the naturalness of accenting these verbs. It is also intriguing that the range of scores appears generally more compressed for these trials

Item group	Accent status	Object status	New-Inference	New-Repeated	Inference-Repeated
Entailment	Accented	New	n.s.	n.s.	n.s.
Entailment	Accented	Old	n.s.	***	***
Entailment	Deaccented	New	n.s.	n.s.	n.s.
Entailment	Deaccented	Old	n.s.	**	marg.
Bridging	Accented	New	n.s.	n.s.	n.s.
Bridging	Accented	Old	n.s.	***	***
Bridging	Deaccented	All data	n.s.	marg.	n.s.

Table 7.9: Experiment 13: Paired comparisons for verb status analysis. n.s. = not significant ($p > .1$). marg. = marginal ($p < .1$). ** = $p < .01$. *** = $p < .001$.

Item group	Verb status	Accent	Object status	Interaction
Entailment	New	***	*	n.s.
Entailment	Inference	n.s.	*	*
Entailment	Repeated	*	***	***
Bridging	New	**	***	*
Bridging	Inference	*	***	n.s.
Bridging	Repeated	**	***	*

Table 7.10: Experiment 13: Summary of models by inference relation and verb relation for accent status comparison. n.s. = not significant ($p > .1$). * = $p < .05$. ** = $p < .01$. *** = $p < .001$.

than in Experiment 11, which is discussed at more length below.

The primary conditions of interest are the old-object trials with deaccented verbs. Crucially, the naturalness difference between deaccented repeated and inference verbs was neutralized in this experiment, in line with the predictions of the accommodation account of deaccenting under nonidentity. However, this finding is difficult to interpret in light of the fact that there was not a clear separation in naturalness ratings between repeated and inferable verbs on the one hand and new verbs on the other. In the entailment items, there was no significant effect of verb status, although in the old-object conditions, there is a numeric trend toward lower ratings with a deaccented new verb. In the bridging items, deaccented new and repeated verbs were rated significantly differently with an old object, but the ratings for inference verbs were intermediate and not significantly different from either of these.

This finding may be driven by the generally middling ratings provided in Experiment 12,

Item group	Verb status	Object status	Accented-Deaccented
Entailment	New	All data	not tested (interaction n.s.)
Entailment	Inference	New	n.s.
Entailment	Inference	Old	*
Entailment	Repeated	New	n.s.
Entailment	Repeated	Old	***
Bridging	New	New	n.s.
Bridging	New	Old	***
Bridging	Inference	All data	not tested (interaction n.s.)
Bridging	Repeated	New	n.s.
Bridging	Repeated	Old	***

Table 7.11: Experiment 13: Paired comparisons for accent status analysis. n.s. = not significant ($p > .1$). * = $p < .05$. *** = $p < .001$.

as mentioned above. Since the range of scores for old-object items is quite compressed, the scores even for canonically “good” sentences, like those with deaccented repeated verbs, reach only the middle of the 7-point scale at best. Since the floor for mean ratings was around 3 in both experiments, it is not surprising that the ratings for conditions with means around 4 or lower were not significantly different from canonically “bad” trials, such as sentences with deaccented new verbs.

For old-object trials, the by-accent comparison showed the canonically expected results for new and repeated verbs, with accentuation reliably preferred for new verbs and deaccentuation preferred for repeated verbs. For inference verbs, the results were mixed. There was no preference between accentuation and deaccentuation in the entailment items, whereas accentuation was significantly preferred in the bridging items, as in Experiment 11.

As mentioned above, the old-object ratings in this experiment, including those for accented verbs, appear to have collapsed toward the middle of the rating scale compared to the results from Experiment 11. Thus, although the difference between new and inference verbs on the one hand and repeated verbs on the other hand remains highly reliable in this experiment, the ratings difference is strikingly smaller in the current experiment than in Experiment 11.

There are several possible explanations for this finding. One is that the addition of the

context sentence itself underlies the middling ratings, with participants having highly eroded intuitions about what prosodic realizations are appropriate for more articulated discourse contexts compared to out-of-the-blue utterances of the type rated in Experiment 11. Another possibility is that this is an effect of removing the “warning” screen before participants heard the sentence; recall that in Experiment 11, participants saw a text representation of the critical sentence in advance, whereas in Experiment 12, this was replaced with the context sentence. It is possible that the participants gave more extreme ratings when they knew the lexical content of the sentence in advance compared to when they needed to process the sentence for the first time in the auditory modality. This was part of the motivation for replicating the study with Experiment 13 including a preview of the target sentence.

Experiment 13: There were relatively few qualitative differences in the results for Experiment 13 compared to Experiment 12. In the verb status comparison, there was now a significant accent-object interaction for deaccented entailment trials. This suggests that the scores were somewhat less clustered than in Experiment 12, with greater differentiation in the effect of verb status in old-object compared to new-object conditions. For deaccented bridging items, the difference between new and repeated verbs was no longer significant.

In the accent status comparison, there was a new verb-object interactions for the new bridging, with accentuation preferred over deaccentuation with an old object. This is not surprising, as new verbs are canonically expected to be accented; the non-significant interaction in the previous experiment may have been due to noise.

There was also a new significant interaction for inferable entailment trials, with a significant preference for accentuation in old-object trials at the $p=.05$ level. The significance level for the preference for accentuation in inferable bridging trials was now $p=.05$ rather than $p=.01$. Finally, the Experiment 12 marginal accent effect on new-object, repeated-verb bridging trials did not persist.

Synthesis and implications: Considered together, the results of the two experiments suggest that the introduction of context slightly improved participants’ assessments of deac-

cented inferable material. Perhaps the clearest evidence comes from the comparison by accent status. In Experiment 11, there was always a highly reliable preference for accenting inferable verbs, with an effect significant at the $p=.001$ level for both entailment and bridging items (and preferences at the same significance level for accenting new verbs and deaccenting repeated verbs). In Experiments 12 and 13, accenting inferable verbs was always numerically preferred, but the preference was impressionistically weaker, with the significance level ranging from not significant in the case of the Experiment 12 entailment trials to $p=.01$ for the Experiment 12 bridging trials. This effect was unique to the inference conditions, as both experiments showed highly reliable preferences ($p=.001$ level) for accenting new verbs and deaccenting repeated verbs, with the exception of the lower significance level ($p=.05$) in the Experiment 12 new-verb bridging trials.

Notably, this improvement for deaccented inferable verbs did not come at the expense of ratings for accented inferable verbs, which were generally still rated comparably to accented new verbs (particularly in Experiment 13). This is not surprising, since it would certainly be felicitous to use contrastive accent on an inferable target if the speaker wanted to highlight a small difference in the meanings of the two verbs.

The findings of the accent status comparison suggest a small amelioration effect of the discourse contexts used here on the naturalness of deaccented inferable material in perception. However, accentuation was still numerically preferred across the board, and often was reliably preferred in statistical terms. Thus, it is not clear that this experiment provides reliable evidence for an effect of context on the acceptability of deaccenting licensing on inferable constituents. The next experiment continues the investigation of the interaction of context and prosody with more conclusive results.

7.3 Experiment 14: Perception of emphasis with presupposition triggers

Experiment 14 replicates Experiment 11 with one change: the presupposition trigger *too* was added to the end of each sentence. The additive particle *too* is known to introduce certain presuppositions to the discourse that are potentially relevant to deaccenting. While the exact mechanisms are complex and a subject of continuing discussion, one persistent characterization is that *too* associates with a focused constituent (Rooth, 1985, 1992b) or a contrastive topic (Krifka, 1998). This association introduces a presupposition that a focus alternative to the clause with the associated element replaced by an existentially closed variable is available in the discourse. On a discourse-driven view, *too* can be analyzed as indicating that a particular piece of information extends a previous contribution to the discourse, e.g., a previous answer to a question under discussion (Beaver and Clark, 2008).

It is clear how this is relevant for deaccenting under inference. The modified stimulus in (132) is an illustrative example.

(132) Ethan astounded Amy, and NAN *surprised Amy*, too.

Here, *too* associates with the subject of the second clause, NAN. This triggers a presupposition that an alternative of the form *x surprised Amy* is available. While there is not an antecedent that has this syntactic form, the meaning is clearly available in the discourse, as *Ethan astounded Amy* entails that *Ethan surprised Amy*. Since calculating the connection between the overt antecedent and this alternative is necessary to verify that the presupposition of *too* is satisfied, by hypothesis, it may also help listeners understand that deaccenting of *surprised* in the second clause is also valid because *astounded* makes the meaning given. This may occur because listeners are more willing to accommodate an alternative antecedent (accommodation account), or simply because they are made more aware of the inferencing relation (grammatical account).

In summary, the experiment investigates whether adding a presupposition trigger like

too can improve naturalness ratings for deaccented inferable constituents, which were determined in the previous studies to be somewhat infelicitous in both production and perception compared to deaccented repeated material. This will be determined by qualitatively comparing the results to those from Experiment 11, which had an identical setup, except that the stimuli did not include presupposition triggers.

7.3.1 *Design, materials, and procedure*

The design, materials, and procedure for Experiment 14 were identical to those for Experiment 11 with one major exception: every stimulus sentence from Experiment 11 was altered to have the word *too* at the end of the sentence (i.e., after the second-clause object).

This change necessitated re-recording the experimental stimuli, as none of the recordings used in the previous perception studies included *too*. The female speaker who recorded the expanded set of stimuli for Experiment 11 returned to record all 48 items in all six conditions. The male speaker was no longer available, so the gender manipulation was eliminated and participants only rated recordings with the female voice. The experimental lists were adjusted so that participants heard an equal distribution of stimuli from each condition after the gender alternation was eliminated.

The female speaker recorded the stimuli in four experimental sessions, as for Experiment 11 (where she recorded one session as a participant in Experiment 9 and three sessions separately). Due to a global health crisis, it was not possible for the speaker to record the stimuli in a sound booth, so she recorded the stimuli remotely and transmitted them for use in the experiment. While the stimuli were not recorded in a sound booth, the speaker had access to high quality sound equipment, and the stimuli were recorded with an external microphone and, impressionistically, featured minimal background noise.

Since the stimuli were recorded remotely, the texts were presented over Ibex Farm instead of PsychoPy. The music phase from the Experiment 9 paradigm was eliminated, as the speaker was able to take breaks at any time during recording. Since Experiment 14 would no

longer feature an alternation between male and female voices, the speaker also recorded new practice and filler stimuli so that her voice would be used throughout the experiment. The critical recordings in the old-object conditions were impressionistically reviewed to verify that they included canonical prosody in the second clause (verb accentuation when new, deaccentuation when repeated) and the speaker was asked to re-record some of the stimuli during the filler recording that did not conform to prosodic expectations.

7.3.2 *Participants*

140 participants took part in Experiment 14, of whom 46 were female and 94 were male. The mean age of the participants was 37.0 years. No participants failed to self-identity as native speakers of English, but the data from 23 participants was excluded from analysis because they had a mean reaction time of less than 1000 ms, indicating that they were not carefully considering their responses. For the participants whose data was analyzed, the mean completion time for the Ibex Farm portion of the experiment was 8 minutes, 44 seconds. Participants received USD 1.00 through the AMT platform for participating.

7.3.3 *Predictions*

The main question of interest in the experiment is whether the addition of *too* can improve naturalness ratings for deaccented inferable constituents. Thus, the results of this experiment will be qualitatively compared to those for Experiment 11. If the ratings for deaccented inferable material improves, it suggests that the empirical generalization from the literature that inferable constituents can be deaccented is not fundamentally flawed, but that lexical inferencing relations alone are not sufficient to license deaccenting. Rather, stronger cues for identity of situational meaning, such as presupposition triggers, are necessary.

If the ratings for deaccented inferable material do not improve, it may point to a deeper flaw in the empirical generalization as described in the literature. Further investigation would be necessary to determine why

7.3.4 Results

The results of Experiment 14, highlighting comparison by verb status, are shown in Figure 7.16 for entailment items and Figure 7.17 for bridging items. Figure 7.18 highlights comparison by accent status for the entailment items, and Figure 7.19 for the bridging items.

A visual inspection of the plots suggests that introducing a presupposition trigger had a strong ameliorative effect on ratings across the board. While in the previous perception experiments, the ratings in “infelicitous” conditions (e.g., accented repeated verbs) could be as low as 3 on the 7-point scale, even the lowest-rated conditions in this experiment have mean ratings above 4.5.

Despite the high ratings across the board, however, there are some telling differences by condition that suggest that adding *too* improved the ratings of deaccented inferable material beyond just the general amelioration effect. In deaccented, old-object conditions, inferable verbs now appear to pattern with repeated, rather than new, verbs. Repeated and inferable verbs received nearly identical ratings in both the entailment and bridging items, while the ratings for new verbs are somewhat lower. However, when the verb was accented, inference verbs continued to pattern with new verbs in receiving slightly higher ratings than repeated verbs. As discussed above, this is not surprising, since accenting an inferable verb can signal that it should be interpreted as contrasting with the antecedent rather than being given.

Another notable trend in the verb status comparison is that new-verb trials appear to have received systematically lower ratings in the new-object conditions, particularly for the entailment items. This may be because these conditions include *too* after a second clause consisting of a new subject, a new verb, and a new object. The presupposition of *too* is difficult to construe as being satisfied here, so there may have been a negative impact on the ratings even though the participants were instructed to focus on the prosody of the sentence. (Indeed, it is not necessarily clear what the “correct” prosody should be for a sentence that may be infelicitous in all contexts.) Since prenuclear position has not been fertile ground for

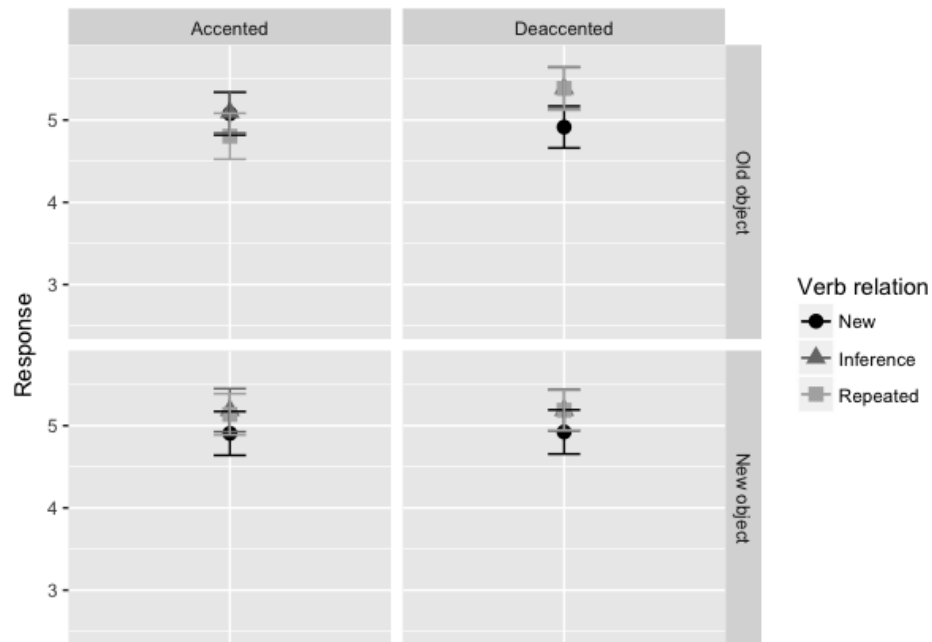


Figure 7.16: Experiment 14 results by verb status, entailment items. Error bars: 95% confidence interval.

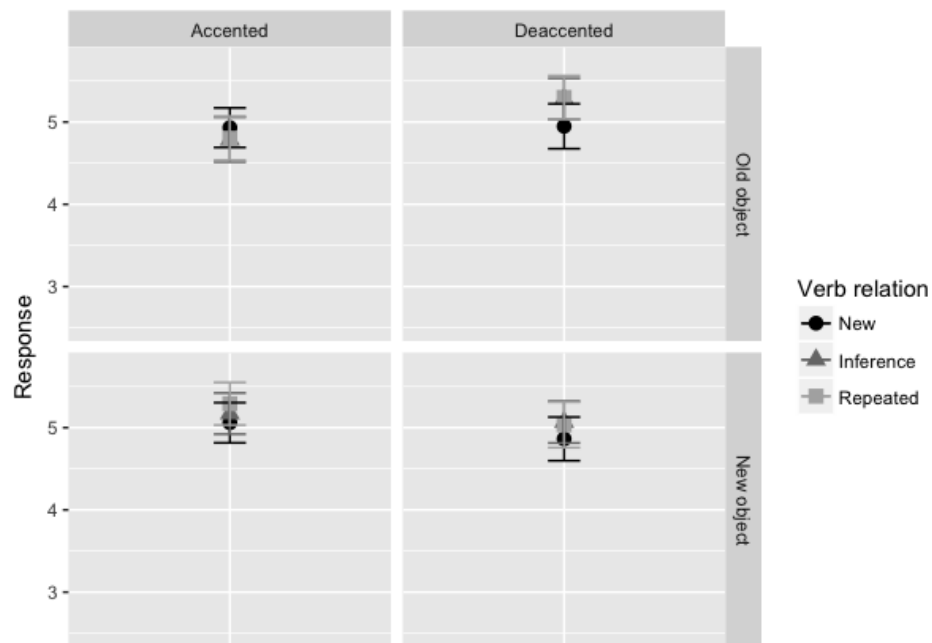


Figure 7.17: Experiment 14 results by verb status, implicational bridging items. Error bars: 95% confidence interval.

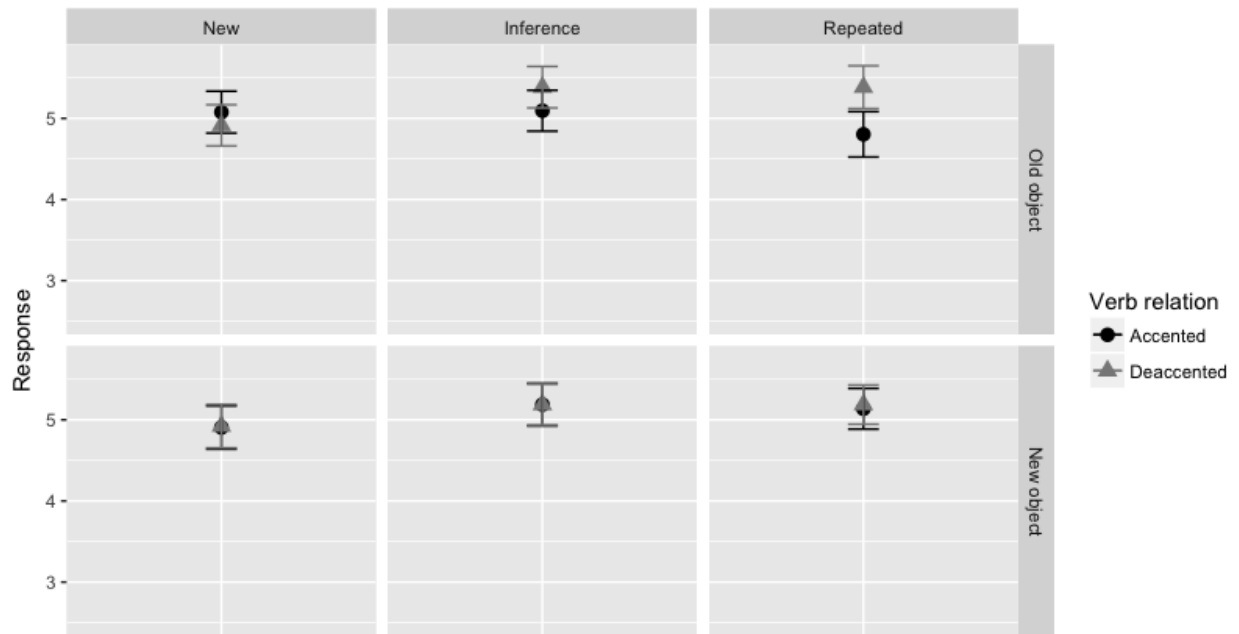


Figure 7.18: Experiment 14 results by accent status, entailment items. Error bars: 95% confidence interval.

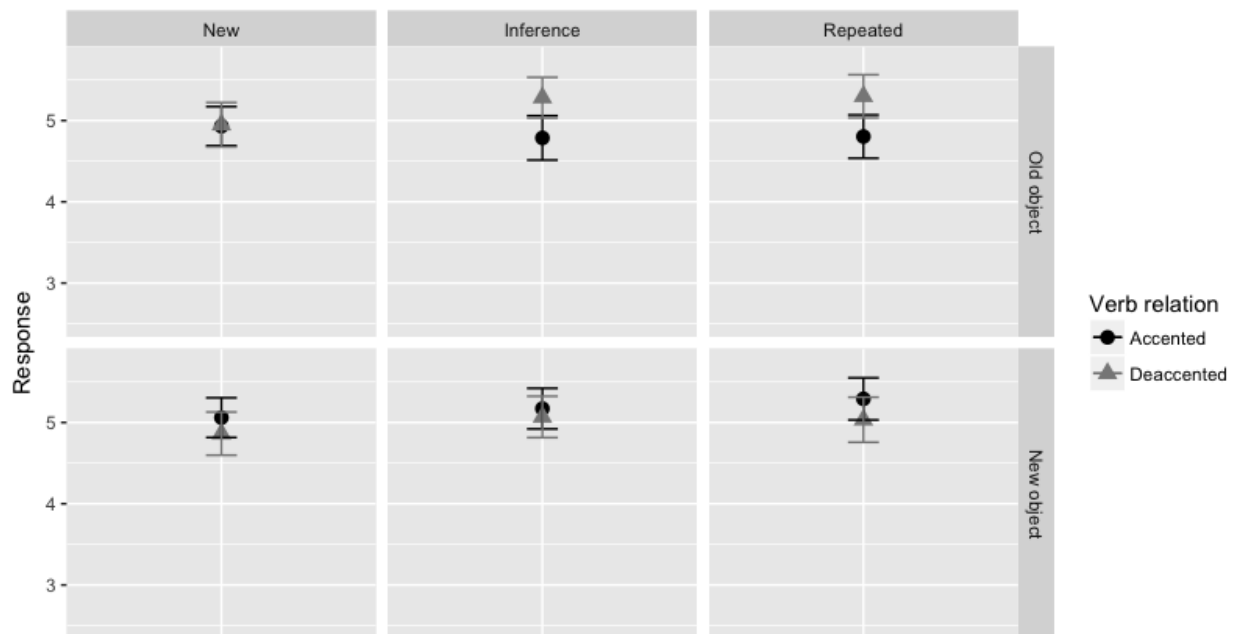


Figure 7.19: Experiment 14 results by accent status, implicational bridging items. Error bars: 95% confidence interval.

useful experimental results about the deaccenting mechanism, this will not be considered in much detail.

Turning to the accent status comparison in the old-object conditions, there appears to be a reliable preference for deaccenting repeated verbs. Interestingly, the preference for accenting new verbs appears to have eroded; there is a numeric trend toward preferring accentuation in the entailment items, while accented and deaccented new verbs received nearly identical ratings in the bridging items.

In the critical inferable conditions, there is now a numeric trend toward a preference for deaccentuation of inferable verbs. The trend appears to be particularly strong in the bridging items, while the results for the entailment items are somewhat more mixed. The trend appears telling compared to the previous perception studies, where accentuation was reliably preferred in out-of-the-blue contexts and numerically or significantly preferred with supportive contexts.

7.3.5 *Analysis*

As in the previous perception studies, analysis began with the construction of logistic mixed-effects regression models for the entailment and bridging data separately. The models contained three-way interactions of verb status, object status, and accent, plus the full component structure of this interaction and random intercepts for participant and item. Unlike in previous experiments, the three-way interaction was not significant in either model (entailment: $p < .1$; bridging: $p > .1$). This may be because the ratings for the new-object observations are much closer to the old-object observations.

Based on the reliably different behavior between new- and old-object conditions in each of the prior experiments, the data were divided by object status for further analysis. Linear mixed-effects regression models were constructed for each inference relation/object status combination with an interaction of verb status and accent, the component main effects, and

random intercepts for participant and item.⁴

For new-object entailment items, there was a significant main effect of verb status ($p < .05$), while the main effect of accent ($p > .7$) and the interaction ($p > .9$) were not significant. Paired comparisons indicated that new-verb sentences received different ratings from inference-verb and from repeated-verb sentences (p 's $< .05$), while these latter did not receive significantly different ratings ($p > .9$).

For new-object bridging items, there was a significant main effect of accent ($p < .05$), while the main effect of verb status ($p > .1$) and the interaction ($p > .7$) were not significant. Paired comparisons were not carried out because there were only two levels for accent; visual inspection of Figure 7.17 indicates that sentences with accented verbs received higher ratings.

For old-object entailment items, there was a significant interaction of verb status and accent ($p < .01$), as well as a significant main effect of accent ($p < .01$) and a marginal main effect of verb status ($p < .1$). Paired comparisons by verb status and by accent were carried out using estimated marginal means. With a deaccented verb, new- and inferable- ($p < .01$) and new- and repeated-verb sentences ($p < .01$) received significantly different ratings, while inferable- and repeated-verb sentences did not ($p > .9$). No pair of verb status conditions received significantly different ratings with an accented verb (p 's $> .1$). Accented- and deaccented-verb sentences received significantly different ratings when the verb was inferable ($p < .05$) or repeated ($p < .001$), but not when it was new ($p > .2$).

For old-object bridging items, the interaction of verb status and accent was significant ($p < .05$). The main effect of accent was significant ($p < .001$), while the main effect of verb status was not ($p > .5$). In the deaccented-verb conditions, new- and inferable-verb sentences received marginally different ratings ($p < .1$). New- and repeated-verb sentences received significantly different ratings ($p < .05$). The ratings for inferable- and repeated-verb sentences were not significantly different ($p > .9$). When the verb was accented, no pair of verb status conditions received significantly different ratings (p 's $> .5$). Accented- and deaccented-verb

4. `lmer` model specification: $\text{Response} \sim \text{Verb} * \text{Accent} + (1 \mid \text{Participant}) + (1 \mid \text{Item})$.

sentences received significantly different ratings when the verb was inferable ($p < .01$) or repeated ($p < .001$), but not when it was new ($p > .8$).

7.3.6 Discussion

The most important finding of Experiment 14 is that introducing the presupposition trigger *too* induced a significant preference for deaccenting inferable verbs in nuclear position. This is in contrast to the previous perception studies, where there was generally a reliable preference for accenting inferable verbs, and when the preference was not reliable, there was still a numeric trend in favor of accentuation.

This finding suggests that deaccenting of inferable constituents can be made acceptable, but that the distribution of such deaccenting is somewhat more restricted than the prior literature has suggested. Previous analysis has tended to appeal overtly to the inferencing relations linking lexical items or phrases as the primary mechanism for licensing deaccenting; if a target was “given” with respect to an antecedent, it could be deaccented. Experiment 11 suggests that lexical inferencing relations on their own are not sufficient. All of the target inference verbs in the current paradigm are highly lexically inferable from their antecedents, yet were not marked as acceptable when deaccented. Rather, Experiment 14 suggests that this type of deaccenting operation requires additional support from the broader context. Presupposition triggers appear to have provided such support, forcing participants to calculate whether the presupposition of *too* was satisfied in such a way that also highlighted the inferencing relation linking the antecedent and target verbs.

A second finding of the experiment was that adding *too* increased ratings of the critical sentences across the board, without respect to verb status, object status, or verb accent. This may suggest that presupposition triggers are very powerful in terms of the types of prosodic encodings they can “rescue.” One puzzling instance is the case of new-verb sentences with old objects (verb in nuclear position). While the previous perception studies generally showed a reliable preference for accenting new verbs in nuclear position, that preference eroded

to indifference between accentuation and deaccentuation in the presence of *too* for both entailment and bridging items.

New verbs canonically would be expected to be accented, and in particular, the new verbs used here received very low ratings of their inferability from their antecedents (below 3 on a 7-point scale) in Experiment 8. Nevertheless, sentences deaccented new verbs were actually rated as fairly felicitous, and roughly on par with sentences with accented new verbs. One interpretation of this is that adding *too* even coerced listeners to read new verbs and their antecedents as “pragmatically identical” in some way even though they were generally quite different in meaning. However, there may be alternative explanations, depending on the precise presuppositions introduced by *too*, that do not require identical readings of new verbs and their antecedents.

As a side note, it has actually been argued that additive particles like *too* may be obligatory when their presuppositions are satisfied in the discourse. Abrusán (2014) gives the example in (133a), which is very similar to the stimuli in the current experimental paradigm, and marks the alternative without *too* as pragmatically infelicitous.

- (133) a. John ate pizza, and Mary ate pizza, too.
b. #John ate pizza, and Mary ate pizza.

One explanation of this is that *too* is a presuppositionally stronger alternative to $\emptyset$, and that use of *too* when felicitous is made mandatory by the maxim Maximize Presupposition (Heim, 1991; Percus, 2006). Another approach holds that uttering the first clause generates the implicature that the speaker is not in a position to give a more complete answer to the question implicit in the discourse (Krifka, 1998; Sæbø, 2004). Using the second clause alone would conflict with this implicature and result in infelicitousness, but adding *too* can discharge the implicature and eliminate this conflict.

The possibility that sentences like (133b), without *too*, are infelicitous calls for cautious interpretation of the results of the prior experiments. Participants in both the production and perception studies may have exhibited unusual behavior if they found the sentences

themselves odd. On the other hand, the results of all of the studies without *too* were highly internally consistent, and the results for new- and repeated-verb conditions conformed to *a priori* expectations for pronunciation in production and naturalness in perception despite the absence of *too*. Nevertheless, it is clear that further investigation of the interaction between prosodic givenness marking and presupposition triggers is warranted, and additional experimentation may reveal whether it is correct to draw strong conclusions about the nature of the mechanisms licensing deaccenting in the absence of presupposition triggers.

7.4 General discussion

The results of the perception experiments presented in this chapter have helped to refine the empirical status of deaccenting of constituents that are inferable from, but not identical to, their linguistic antecedents. The prior literature has suggested that antecedent-nonidentical material is deaccentable when it is reasonably marked as given due to a number of possible relations with an antecedent: when it co-refers with an antecedent, it is entailed by an antecedent, modulo existential closure, or it is reasonably concluded from the antecedent on the basis of broader world knowledge.

The production study from the previous chapter, Experiment 9, and the perception study examining prosody in out-of-the-blue contexts, Experiment 11, called this claim into question. The verbs used in the “inference” conditions of these experiments were rated as highly inferable on the basis of their linguistic antecedents. Nevertheless, in nuclear position, speakers did not deaccent these verbs in production, and listeners gave them only middling acceptability ratings (reliably lower than repeated verbs) when they were deaccented.

Despite this initial challenge, the subsequent perception studies indicated that deaccented inferable material can be made acceptable by support in the broader discourse context. Experiments 12 and 13 added a discourse context supporting a pragmatically identical reading of inferable verbs and their antecedents. The result was that what was previously a strongly reliable preference for accenting inferable verbs degraded to a numeric preference for ac-

centing that sat closer to the cusp of statistical significance. Experiment 14 added the presupposition trigger *too* to the end of the stimuli, which resulted in a significant preference for deaccenting inferable verbs.

These results indicate that deaccenting of inferable material is possible, but is less restricted than has been presented in the prior literature. Grammatical accounts of this phenomenon predict that deaccenting is licensed and mandatory whenever it is apparent that the target is inferable. This most clearly applies to entailed targets, where the inferencing relation should be obvious to both speakers and hearers. The accommodation account of deaccenting predicts that deaccented nonidentical constituents are accommodation-seeking material and should straightforwardly trigger accommodation of an alternative antecedent, within reason. The results from this chapter suggest that these operations are not sufficient. The grammatical account's prediction that lexical inferability is sufficient to license deaccenting was not borne out, nor was the accommodation account's prediction that mere deaccenting of a nonidentical word could trigger accommodation.

Instead, the results indicate that deaccenting of nonidentical material requires support from the discourse context other than the antecedent and the target. The felicitousness of deaccented inferable material appeared to increase as a function of contextual support for identical readings of the antecedent and target. When there was a moderately supportive discourse context (Experiments 12 and 13), the preference for accentuation was weakened, and when the additive particle *too* introduced a formal presupposition that the target verb's meaning was available in the discourse, the preference reversed into an advantage for deaccenting.

The finding that deaccenting of inferable constituents is not licensed by lexical inferencing relations alone, but increases as a function of contextual support, is problematic for the grammatical account of licensing. This account gives a uniform mechanism for licensing canonical deaccenting under identity and deaccenting of inferable material. As such, it leaves little room for differences in felicitousness between the two. The most reasonable

path for maintaining such an account is that speakers and hearers in the experiments where deaccenting inferable material was not observed were not aware of the inferencing relations linking the critical targets and their antecedents. This proposal meshes well with the observation that felicitousness increases as contextual support increases, since contextual support is likely to increase participants' awareness of the inferencing relations. However, it is difficult to reconcile with the norming results of Experiment 8, where participants gave some of the most extreme Likert ratings observed in the entire dissertation indicating that the critical target verbs were inferable from their antecedents.

By contrast, the results are readily compatible with the accommodation account of deaccenting licensing under nonidentity, although the empirical distribution of such deaccenting is slightly more restricted than has been proposed in the accommodation literature. The accommodation account schematizes deaccenting as ungrammatical in cases where there is not string identity with an antecedent. As such, the model gives a clear theoretical path for felicitousness differences between deaccented repeated material and deaccented inferable material. In addition, the accommodation account interfaces well with the observation that deaccenting felicitousness sits on a cline in response to contextual support. It seems natural that additional contextual support can help listeners conclude that accommodation of the alternative antecedent needed to license deaccenting is reasonable. The only modification that is necessary to the accommodation account is that the notion of accommodation-seeking material presented in the literature was somewhat too permissive; a deaccented constituent on its own, even when highly inferable, is evidently not sufficient to trigger accommodation.

One question that remains is why there is such a sharp disconnect between the felicitousness of deaccenting observed in the experiments here and its treatment in introspection in the literature. For the most part, deaccenting of inferable constituents is marked as straightforwardly acceptable in the literature, and it is treated as a core phenomenon for which the theory of deaccenting must account. One possibility is that, despite the theoretical formulation that mere inferability is sufficient to license deaccenting, the literature examples all

contain support for deaccenting in the context that goes beyond merely the lexical inferencing relations linking the antecedent and target. This might include co-reference or presupposition triggers such as *too*, but also more informal support. For example, in the *Republican* example, repeated in (134), *First* and *Then* might help construe the events described in the first and second clause as situationally identical to a degree beyond what the listener would conclude out-of-the-blue, even if they believe that the inferencing relation holds.

- (134) First, John called Mary a Republican. Then, BILL *insulted Mary*.
(Lakoff, 1968)

A second possibility is that consideration of such examples in the literature includes additional implicit support for deaccenting, although this support is not made explicit in the example itself. Examples with deaccented inferable material are often presented in conjunction with additional text explaining why the example “works” – for instance, outlining the connection between calling someone a Republican and insulting them in (134). This may suggest that these examples do not work well out of the blue, but rather require the type of contextual support explored in the experiments here. This is not to say that the examples cited are not felicitous, but rather, that they only work well when embedded in a supportive discourse. Presenting only the critical sentence as an example leads to the impression that the example should work well on its own, but it actually takes additional text discussion to create the conditions that license deaccenting comparably to a supportive naturalistic discourse. Future work in the vein of Experiments 12 and 13 examining the effect of the broader context may help provide a more detailed account of the interaction between the discourse at large and the prosodic encoding of inferable material.

Yet another possibility is that the nature of the experimental tasks limited participants’ impressions of the felicitousness of deaccenting of inferable material relative to what would happen in a naturalistic discourse. Since the participants were not actually engaging in a discourse with a real interlocutor, they may not have devoted sufficient attention to the relations that license deaccenting under nonidentity, such as entailment. It may be the

case that listeners only truly attend to the discourse at the level required to naturalistically calculate givenness in the way described in the literature when there are real “stakes,” for instance the chance of miscommunication. In the experiments, since there are no such stakes, participants may have readily calculated that repeated words were given, since this requires relatively less attention, while failing to mark entailed and other inferable material as given. In other words, participants may not have constructed a fully realized model of the discourse in the experiments in the same way they would in real communication because of the lack of an interlocutor. Thus, future research might focus on evaluating deaccenting in a more ecologically valid paradigm (e.g., a study with a confederate) to determine if the acceptability of deaccenting under nonidentity increases in face-to-face communication.

One surprising result of this chapter was the powerful ameliorative effect of introducing the additive particle *too* in Experiment 14. It is relatively clear how *too* promoted deaccenting of inferable material, since it forces listeners to calculate whether the presupposition that the target verb’s meaning is given in the discourse is satisfied. However, adding *too* also yielded very high ratings for configurations that would canonically be marked as unacceptable, such as accented repeated verbs. There was still a reliable preference for deaccenting inferable verbs rather than accenting them with *too*, but it is worth disentangling the extent to which the acceptability of deaccenting inferable material with *too* is due to the processes outlined in the accommodation account rather than to this powerful amelioration effect. This and additional future directions of research are discussed in the conclusion that follows.

CHAPTER 8

CONCLUSION

The final chapter of this dissertation considers the results for verb phrase ellipsis and deaccenting together. The sections below first summarize the prior debates and experimental findings from each of the dissertation’s halves in turn. Following this, the final section discusses the implications of these findings for the prospects of a unified analysis of the interaction between linguistic antecedents and broader context information for verb phrase ellipsis and deaccenting, as well as broader implications for the theory of anaphora and context-dependent constructions. This final section also highlights some possible directions for future work on these and other constructions that may further improve the theoretical understanding of these mechanisms.

8.1 Summary of findings: Verb phrase ellipsis

The first half of the dissertation focused on the mechanisms underlying the interpretation of verb phrase ellipsis in rich discourse contexts. Chapter 2 discussed the relevant prior literature on ellipsis. A major debate in this literature concerns the nature of the relationship between ellipsis sites and their antecedents. One possibility is that there is a requirement for syntactic isomorphism between the two. In general, this predicts that formal mismatches between ellipsis clauses and their linguistic antecedent should be unacceptable, but various methods have been proposed to “rescue” certain ungrammatical sentences that are judged acceptable. A second possibility is that ellipsis sites rely on the presence of a semantically isomorphic meaning that is available in the discourse. This predicts that formal mismatches between ellipsis sites and antecedents should be permitted, and various mechanisms have been proposed to explain why some mismatch sentences are judged unacceptable.

A potentially relevant piece of data that has received relatively little prior attention is the case of antecedentless or exophoric ellipsis. Relatively recent corpus investigation has

indicated that verb phrase ellipsis can be used in the absence of a linguistic antecedent to express non-idiomatic meanings, although the distribution of such uses is quite limited. At first blush, this finding appears to support semantic accounts of the ellipsis identity relation, as the fact that such ellipsis sites can be interpreted at all suggests that interpretation involves finding a salient VP-meaning in the discourse at large. However, there are possible interpretive mechanisms for such ellipsis sites that would allow for the maintenance of a syntactic identity requirement, such as Merchant's (2004) proposal that comprehenders understand the elided material in exophoric cases to be the deep verbal anaphor *do that*.

This dissertation focused on an empirical gap inspired by recent investigation of exophoric ellipsis. The bulk of the literature on VPE focuses on exclusively linguistic discourses, testing the complex interactions between ellipsis sites and various linguistic antecedents. Likewise, work on antecedentless ellipsis has focused on the question of whether VPE can felicitously be used in the absence of a viable linguistic antecedent, and whether such ellipses are interpretable, and so this research has not considered examples with useable linguistic antecedents.

These approaches leave open the question of how VPE is interpreted in "complex discourse contexts" – that is, discourses in which there is both a viable linguistic antecedent VP that could be used to interpret an ellipsis site, but also salient information in the broader discourse context that could potentially be considered during interpretation. If interpretation exclusively considers the content of the linguistic antecedent, it suggests a mechanism of resolution under syntactic identity. If information from the broader context is considered even when there is a viable linguistic antecedent, it suggests that interpretation considers the discourse more holistically, and perhaps primarily deals with finding an antecedent meaning for the ellipsis site.

The end of Chapter 2 outlined some possible models for context-situated VPE interpretation, which were noted to sit on a spectrum from purely syntactic to purely semantic. At the syntactic end of the spectrum is a model requiring strict syntactic isomorphism with

a linguistic antecedent during interpretation. Since a linguistic antecedent is required for interpretation, a separate mechanism is required for interpreting exophoric instances, such as Merchant’s elided *do that* proposal.

A more relaxed syntactic account would posit interpretation under identity with a linguistic antecedent, but with the provision that covert alternative antecedents could be constructed under certain circumstances. Such a mechanism has precursors in prior accommodation (Fox, 2000; van Craenenbroeck, 2012; Thoms, 2015) and repair (Arregui et al., 2006) accounts of acceptable mismatches. However, in complex discourse contexts with viable (matching) linguistic antecedents, the exact circumstances under which alternative antecedents would be constructed is not immediately apparent, as these prior accounts treated formal antecedent-ellipsis mismatches as the trigger for the operation.

At the far semantic end of the continuum is a purely discourse-driven model with a “simple” set of heuristics for identifying the antecedent meaning. This account is inspired by Miller and Pullum’s (2013) and Miller and Hemforth’s (2014) account of VPE licensing. According to this proposal, the felicitous use of VPE requires that certain types of alternative sets are made available in the discourse. Interpretation involves finding the alternative set (or one of the sets) that licensed the ellipsis and selecting the member of that set that is compatible with the ellipsis clause. Importantly, listeners consider all of the information available in the discourse context, both linguistic and nonlinguistic, when interpreting ellipsis sites. One interpretation of this – though not one explicitly appealed to by the authors – is that listeners categorically access the alternative set that is most salient in the discourse, or give consideration to particular alternative sets in direct proportion to their salience in the discourse. In other words, there is a one-to-one mapping between a particular proposition’s salience in the discourse at large and its availability under ellipsis.

A more permissive discourse-driven account would appeal to a more articulated set of heuristics for identifying an antecedent meaning. Hardt (1993), arguing for a semantic identity model of VPE, sketches some heuristics that might govern the selection of antecedent

meanings in exclusively linguistic discourses, including preferences for subject coreference, special clausal relations, and recency. Such a mechanism is especially appealing for approaches that treat VP ellipsis sites as unpronounced anaphors, as there is potential for the interpretation heuristics to mirror the form-specific biases in interpretation that are known to be in effect for other anaphors, like pronouns (Kehler and Rohde, 2013).

Chapter 3 covered the experimental investigation of VPE interpretation in complex discourse contexts. The experimental design probed whether participants included number information in their interpretation of target verb phrases. The target sentences occurred with both local linguistic antecedents and comic strips representing a broader discourse context (nonlinguistic or mixed linguistic-nonlinguistic). Both types of context varied in the level of support they provided for a number-modified interpretation of the relevant action. Thus, when the target sentence was elliptical or otherwise anaphoric, participants' preferences for including number information in the interpretation of the ellipsis site should indicate whether they were considering information from only the syntactic antecedent or from the discourse at large.

Experiment 1 conducted an initial investigation of VPE interpretation using the experimental paradigm described above. In exophoric cases, ellipsis interpretation largely proceeded according to the information available in the broader discourse context, as expected. When there was a linguistic antecedent that contained number information (a numeral), participants reliably included number information in their interpretation of the ellipsis site, and interpretation was largely unaffected by information in the broader context. Interestingly, when the linguistic antecedent did not contain number information, interpretation was affected by the broader context, with increasing number information outside the antecedent leading to increased consideration of number information in the ellipsis interpretation. This result was argued to be problematic for the strict syntactic identity account, since there was a clear effect of information from beyond the local antecedent VP in ellipsis interpretation.

Experiment 2 tested whether the nonidentity effect observed in Experiment 1 was unique

to ellipsis by testing the interpretation of sentences with fully realized VPs corresponding to the antecedent-supported ellipsis interpretations. While there was still some effect of the broader context manipulation even on the interpretation of full-VP sentences, the effect was smaller in magnitude than in Experiment 1, suggesting that the nonidentity effect is at least partially driven by ellipsis itself.

Experiment 3 assessed the discourse coherence of number-modified and -unmodified sentences in the discourse at large prior to the utterance of the Experiment 1 elliptical sentences. The coherence ratings were intended as a preliminary estimate of the availability of number-modified or -unmodified propositions in the discourse context. The ratings were compared to the results of Experiment 1 to assess whether the “simple” discourse-driven model is an adequate account of VPE interpretation in context. There were substantial mismatches between the coherence ratings in Experiment 3 and the interpretations available under ellipsis in Experiment 1. In particular, in the condition with an unmodified antecedent and salient context, the number-modified and -unmodified sentences were rated as equally coherent in the discourse, but the unmodified (polar) interpretation was significantly preferred under ellipsis. This suggests that the form of the linguistic antecedent constrains the distribution of possible VPE interpretations beyond its simple contribution to salience in the discourse at large, as measured by listener assessments of discourse coherence. This finding was argued to be problematic for the “simple” discourse model, as there was not a straightforward mapping between a proposition’s availability in the discourse at large and its availability under ellipsis. However, it is worth questioning how reliable the coherence ratings were as a measure of discourse availability, a question that was subsequently addressed by a design change in Experiments 5 and 6.

Experiment 4 tested interpretation of the verbal anaphor *do that* in the same paradigm that was used to test VPE in Experiment 1. As expected, *do that* was more sensitive to information from the broad discourse context, with number information considered in the anaphor’s interpretation more readily with support for number information from outside

the local linguistic antecedent. However, it was suggested that VPE and *do that* might lie on a spectrum of sensitivity to broad discourse information rather than on two sides of a categorical surface-deep divide (cf. Hankamer and Sag, 1976). As the previous experiments showed, VPE is sensitive to manipulations in the context outside the local linguistic antecedent, indicating that it is not perfectly “surface.” Conversely, *do that* appears to be constrained by the form of the linguistic antecedent, indicating that its interpretation is guided strictly by simple discourse considerations. (For instance, comparing the results of Experiment 4 to those of Experiment 5 suggests that in the unmodified/salient condition, the scalar interpretation is less available for *do that* than in the discourse at large.) Finally, there were significant differences in the interpretations available for antecedentless *do that* versus VPE, calling into question an analysis of exophoric VPE positing that listeners infer that the elided content is *do that*. However, an analysis positing an elided anaphor with different interpretation preferences, such as *do it*, remains viable.

Considered together, the results of the experiments in Chapter 3 were argued to constrain the hypothesis space of possible models of VPE interpretation from both ends of the spectrum. The results of Experiment 1 showed a sensitivity on the part of VPE to information available in the broad discourse context outside a local VP antecedent, contra the predictions of a strict syntactic account of interpretation with a separate “escape hatch” mechanism for exophoric VPE interpretation. At the same time, there was a large mismatch between the discourse availability of propositions in Experiment 3 and their availability under ellipsis in Experiment 1, which argues against the “simple” discourse model.

Two more nuanced approaches to VPE interpretation were argued to be compatible with the Chapter 3 data. One model is the “complex” discourse approach, according to which VPE interpretation invokes identity with an antecedent meaning, but the heuristics for identifying this meaning go beyond simply accessing propositions based on their relative salience in the discourse. The challenge that remains for this approach is to actually spell out some or all of the heuristics that go into identifying an antecedent. One possibility, from

Miller and Pullum (2013) and Miller and Hemforth (2014), is that local linguistic antecedents are more available to the processor than a representation of the discourse at large. However, more work would be necessary to identify and test the full set of heuristics that govern the identification of an antecedent meaning, taking into account both complex discourse contexts like the ones studied here, but also more complicated linguistic discourses like those studied by Hardt (1993).

The second possibility is a syntactic identity model with an accommodation or repair mechanism allowing for the construction of covert alternative antecedents. The main challenge for this account is to explain why alternative antecedents would be constructed in the experimental discourses. In the experimental stimuli, besides the exophoric conditions, there was always a viable VP antecedent in the sense that it was formally compatible with the ellipsis clause. By contrast, prior accommodation and repair accounts treat formal mismatches as the trigger for the construction of alternative antecedents. One possibility that was suggested was a noisy channel-inspired approach (Levy, 2008), according to which listeners monitor the discourse at large for possible questions under discussion and consider alternative ways the speaker may have encoded their utterance to achieve their communicative goals. As for the complex discourse model, though, considerable further theorizing and experimentation would be necessary before settling on such a model.

Building on the experimental results from Chapter 3, Chapter 4 sought to find further evidence for the “middle of the road” approaches to context-situated VPE interpretation using a probabilistic modeling methodology. First, Experiments 5 through 7 expanded on the results of Experiments 1 and 3 using a forced-choice methodology. In particular, Experiments 5 and 6 collected forced-choice judgments of whether number-modified or -unmodified propositions were more salient in the discourse at large, while Experiment 7 collected VPE interpretation judgments with a forced choice between a number-modified or -unmodified interpretation. Unlike the Likert-scale responses from Chapter 3, the forced-choice paradigm allowed for modeling the prior discourse and VPE interpretation in terms of the probability

of selecting one a particular proposition or interpretation. The results for VPE interpretation in Experiment 7 were largely comparable to those in Experiment 1. There were some differences in the results for Experiments 5 and 6 compared to Experiment 3, with the coherence measure underestimating the strength of a number-modified proposition in some conditions. However, since the coherence measure already overestimated the strength of a modified proposition compared to ellipsis interpretation in the unmodified/salient condition (where the polar interpretation was significant preferred in Experiment 1), these differences do not undermine the conclusion that the simple discourse approach is not an adequate model of VPE interpretation in context.

The rest of the chapter considered several probabilistic models of VPE interpretation. The models transformed the prior probabilities of saying that a number-modified versus -unmodified proposition was at issue, measured in Experiments 5 and 6, to predicted posterior probabilities of a number-modified or -unmodified interpretation of ellipsis, which was compared to the empirical posteriors measured in Experiment 7. There were three different classes of model. The “identity” models considered only interpretations that would arise if the ellipsis site were interpreted with the antecedent VP *in situ*. The “discourse” models implemented the “simple” discourse strategy of preferring a particular interpretation in proportion to its salience in the discourse at large. The “hybrid” models combined these two strategies in a proportion determined by a free parameter. In addition, each of these models had two instantiations, one with an allowance for interpretation at random from a list of possibilities, and one without this allowance.

Impressionistic comparison of the predicted posteriors according to the maximum-likelihood instantiation of each model with the empirical posteriors from Experiment 7 suggested that either hybrid model provided a substantially better fit to the empirical data than either identity model or either discourse model. This was further confirmed by comparing the models pairwise using Bayes factors, a model comparison technique that abstracts over the possible values of any free parameters rather than examining only the maximum likelihood instantia-

tion. Thus, the modeling analysis in Chapter 4 further indicated that context-situated VPE interpretation is best conceptualized as a “hybrid” of interpretation under linguistic identity and interpretation with respect to the context at large; the salience of competing propositions in the discourse is taken into account, but the content of the linguistic antecedent constrains the distribution of possible interpretations beyond its mere contribution to salience.

8.2 Summary of findings: Deaccenting

The second half of the dissertation focused on the relative roles of linguistic antecedents and inference in anaphoric deaccenting. It has long been recognized that there is a principle relationship between a constituent’s status as “new” or “given” in a particular discourse and its prosodic realization with or without a pitch accent, respectively. However, at least two complications of this observation have made necessary a fairly complicated theoretical architecture to explain the conditions under which a constituent should be accented or deaccented.

The first complication is that deaccenting requires some isomorphism in structural position between a potential deaccenting target and its antecedent. For instance, if a particular proper name is used in subject position, and then again as a direct object in the next clause, it will generally not be felicitous to deaccent the second iteration of the name. This observation has received considerable attention in the literature, and theories of deaccenting licensing are generally well equipped to deal with at least simple cases.

The second complication is the observation that particular constituents can be deaccented in the absence of an identical antecedent, and in some cases, the antecedent for the deaccented material is radically different in form. A canonical example of this phenomenon is deaccenting of the verb *insulted* with the antecedent *called a Republican* (Lakoff, 1972; Rooth, 1992a; Tancredi, 1992). The intuition underlying such examples is that a constituent can be deaccented if its meaning is made available by some antecedent, even if the antecedent is not formally identical to it. In the simplest cases, this requirement is satisfied because the

antecedent entails the target (modulo type shifting). In more complex cases, like the *Republican* example, more complex world knowledge is required to conclude that the meaning of the deaccenting target is available, and there is room for subjective assessment of whether deaccenting is licensed.

There are two distinct approaches to accounting for this second complication in the theory of deaccenting. The first builds inferencing relations directly into the grammar of deaccenting. A prototypical example is Schwarzschild's (1999) requirement that every constituent be GIVEN. Informally, GIVENness means that for a particular target constituent, there is an antecedent that entails the result of replacing any F-marked (and by extension, accented) material in the constituent with existentially bound variables. This means that constituents can be (and, by application of the constraint AVOIDF, must be) deaccented not only when there is an identical constituent in an isomorphic position in an antecedent, but also when they are entailed by a constituent in an isomorphic position. Presumably, this notion of entailment, or a minimal extension of it, also includes more distant cases like the *Republican* example.

The second approach, spelled out perhaps most explicitly by Fox (2000), posits that the grammar of deaccenting actually requires exact identity between a deaccenting target and a linguistic antecedent. Acceptable instances of deaccenting under nonidentity are actually due to accommodation of a covert alternative antecedent. Deaccented material without an identical correlate in the antecedent is *accommodation-seeking material* and triggers construction of alternative antecedents as long as the meanings of those antecedents are sufficiently supported by the discourse.

Cases of deaccenting under nonidentity, like *call a Republican/insult*, are widely cited in the literature on deaccenting and constitute a fundamental piece of data for which theories of deaccenting are meant to account. Despite this, however, these examples, and their status relative to more canonical examples of deaccenting under identity, have received little systematic investigation. The main exception to this is a study by Chodroff and Cole (2019)

investigating deaccenting of noun phrases under identity and nonidentity, which was completed roughly concurrently with the studies presented in this dissertation. Given this empirical gap, the main focus of the second half of this dissertation was to explore the licensing and acceptability of deaccenting for antecedent-identical constituents, inferable constituents, and discourse-new constituents, and to determine whether the results of this investigation speak to the means by which deaccenting under nonidentity should be treated in the theory of deaccenting licensing.

Chapter 6 opened the investigation of deaccenting under nonidentity with a production study, Experiment 9. In this experiment, participants read highly controlled sentences of the form *SVO and SVO*. In these sentences, the second-clause object could be repeated or new relative to the first clause. Critically, the second-clause verb could be new/contrastive, repeated, or could be related to the first-clause verb by one of two inferencing relations. In half the items, the second-clause verb was intended to be entailed (modulo type shifting) by the first verb, while in the other half, the two verbs were linked by an informal bridging relation (e.g., it would be reasonable to accommodate an alternative first clause containing the second-clause verb since the represented actions are highly likely to co-occur). Experiment 8 first verified that the existentially-closed meanings of the inferable-condition verbs were made highly available by the existentially-closed meanings of their antecedents.

The Experiment 9 analysis measured three correlates of prosodic prominence in English – intensity, f_0 , and duration – on all second-clause vocalic nuclei. Analysis of the stressed object nuclei indicated that all three correlates of emphasis showed higher values on new objects compared to repeated objects, confirming that the experimental procedure and analysis was sensitive to independently expected accenting behavior. The analysis of the critical verbs showed that discourse-new verbs were reliably accented, while repeated verbs were reliably deaccented, in accordance with *a priori* expectations. However, inferable verbs in both categories had reliably higher phonetic values than repeated verbs, while they were not reliably different from the values for new verbs.

Experiment 10 tested whether there was a mismatch between the phonetic features analyzed and listeners' holistic assessments of emphasis by playing isolated second-clause recordings from Experiment 9 for naive listeners and asking them to rate individual words as emphasized or not emphasized. The results were qualitatively identical to Experiment 9; new objects were rated as more emphasized than old objects, while new and inferable verbs patterned together as emphasized to the exclusion of repeated verbs, which were not emphasized.

The findings that inferable verbs received roughly identical realizations to new verbs in production conflicts with intuitions from the literature on deaccenting. Proponents of both grammatical and accommodation accounts of deaccenting under nonidentity mark such examples as straightforwardly acceptable, and treat such examples as a core phenomenon to be accounted for by the theory of deaccenting licensing. The observation that inferable verbs were not deaccented in production calls into question the degree of empirical isomorphism between deaccenting under nonidentity and canonical deaccenting, and raises the possibility that deaccenting under nonidentity need not be treated by the core theory of emphasis licensing.

The Chapter 6 results were noted to be more compatible with the predictions of the accommodation account of deaccenting under nonidentity than those of the grammatical account. According to the grammatical account, deaccenting of entailed or otherwise available material is automatically ruled in, and in fact is often characterized as mandatory due to the application of constraints like Schwarzschild's (1999) AVOIDF. On the other hand, the accommodation model treats deaccented nonidentical material as ungrammatical, although it is expected to be marked acceptable in many cases due to the relative ease with which an alternative antecedent can be accommodated.

While the production evidence favored the accommodation account, several caveats were raised at the end of Chapter 6 that were argued to preclude strong conclusions about the nature of the deaccenting licensing mechanism. One important observation is that the speakers

in the production study did not plan their own utterances, but were reading predetermined stimuli. Deaccenting of nonidentical material should send a strong cue about the discourse status of the target constituent, but given that the participants did not have an interlocutor, nor were they invested in the “discourse” in which they were speaking, they may not have recognized that such a prosodic encoding was licensed, or felt that it was necessary if they did realize it was licensed. In other words, naturalistic deaccenting under nonidentity involves the explicit choice to encode the target constituent as given as a means to help the listener make sense of the discourse, whereas participants in the experiment did not have a listener, nor any clear motivation to mark the target verb as given.

A second caveat is that the production participants may not have been aware of the inferencing relations linking the first- and second-clause verbs. While Experiment 8 indicated that the target verbs in the inference conditions were all made highly available by their antecedents, the measurement taken in this experiment may not map straightforwardly onto the extent to which their co-occurrence in a discourse supports deaccenting. Another possibility is that the production participants were only processing the task at a “surface” level, allowing them to recognize the connection between identical constituents and perform the associated deaccenting, but ignore more nuanced lexical relationships such as entailment, leading to accenting where deaccenting might otherwise be licensed.

In response to these considerations, Chapter 7 focused on listeners’ perception of deaccenting under various discourse conditions. A major limitation of the production paradigm was that participants had free choice of prosodic encoding, meaning there was always a possibility of accenting when deaccenting might be licensed. In a perception study, on the other hand, the listener must contend with the stimulus as presented, whether it contains accenting or deaccenting. In other words, listeners must judge whether a particular sentence is felicitous given that the speaker has already chosen to accent or deaccent particular constituents.

The experimental paradigm used throughout Chapter 7 involved cross-splicing first- and

second-clause recordings from the production study paradigm. Verbs that were produced under canonical accenting conditions or canonical deaccenting conditions could thus appear in sentential frames making them discourse-new, repeated, or inferable. Participants were asked to rate how natural the sentences sounded, with a particular emphasis on the “melody” or “tune” of the sentence, which was intended to help naive listeners focus on the prosody of the sentence.

The first experiment of the chapter, Experiment 11, simply played the manipulated recordings for the participants. This was treated as a proxy for the use of such sentences in “out-of-the-blue” contexts. In such contexts, deaccenting under nonidentity is supported only by the lexical inferencing relations linking the antecedent and target verb, whether as a direct component of the grammar (grammatical account) or as support for resolving the presence of accommodation-seeking material (accommodation account). The results of Experiment 11 were comparable to those of the production study. Sentences with accented verbs were rated highly when the verb was new or inferable, and significantly worse when the verb was repeated. Conversely, sentences with deaccented verbs were rated highly when the verb was repeated, and received lower ratings when the verb was new or inferable. New and inferable verbs generally did not receive significantly different ratings.

The result of Experiment 11 was surprising from the perspective of the prior deaccenting literature. Both the grammatical and accommodation accounts treat deaccenting of nonidentical material as straightforwardly acceptable, yet listeners in Experiment 11 reliably treated inferable verbs as though they were discourse new. The rest of the chapter investigated this further by testing whether additional contextual support for the inferencing relation, beyond just the lexical relationship between the antecedent and target verbs, could improve the acceptability of deaccented inferable verbs.

Experiments 12 and 13 situated the manipulated stimuli in a small discourse, with participants reading a context sentence that was attributed to the stimulus speaker before hearing the recorded sentence. The context sentences were constructed to provide moderate support

for a “situationally identical” reading of the antecedent and target verbs – for instance, for the verb pair *hug* and *embrace*, a context noting that several instances of greeting occurred might support the conclusion that the actions described by the two clauses are comparable and the two verbs should not be understood as contrasting.

The results of these two experiments showed that a supportive context improved the acceptability of deaccented inferable material. Whereas in Experiment 11, there was a reliable preference for accenting rather than deaccenting inferable verbs, in the context experiments, this preference eroded, and sentences with accented and deaccented inferable verbs received roughly comparable ratings. At the same time, there was still a reliable preference for accenting new verbs and for deaccenting repeated verbs, indicating that the effect of context was limited to the inference conditions.

Finally, Experiment 14 was a replication of Experiment 11, except that the stimuli were reconstructed so that every sentence had the presupposition trigger *too* at the end of the second clause. The expectation was that the presence of a presupposition trigger indicating similarity between the two clauses might help listeners to construe deaccented inferable material as given. This prediction was borne out in the results. While there was a preference for accenting inferable verbs in Experiment 11 and indifference between accenting and deaccenting in Experiments 12 and 13, there was a reliable preference for deaccenting in Experiment 14. Further, in the comparison by verb status, deaccented repeated and inference verbs patterned together to the exclusion of new verbs, whereas new and inference verbs patterned together in the previous experiments. Interestingly, the addition of the presupposition trigger also led to increased ratings for structures that would canonically be treated as infelicitous, such as sentences with accented repeated verbs. This may indicate that presupposition triggers have a general mitigating effect on otherwise substandard prosodic configurations, and further research may be needed to disentangle this general effect from the improvement in scores that is specific to deaccented inferable material.

In sum, the results of Chapter 7 indicated that the felicitousness of deaccenting under

inference is highly sensitive to the degree of contextual support for the inferencing relation, with ratings sitting on a cline from less acceptable to very acceptable. Despite the fact that all of the inference-condition verbs were highly inferable based on their lexical relationships to their antecedents, listeners reliably preferred for them to be accented in out-of-the-blue sentences. When participants heard the same sentences with a discourse context supporting a situationally identical reading for the antecedent and the inferable verb, there was no preference for accenting versus deaccenting the verb. Finally, in the presence of a presupposition trigger, there was a preference for deaccenting rather than accenting inferable verbs.

Together, the deaccenting results were argued to be more compatible with the accommodation account of deaccenting under nonidentity than with the grammatical account. Under the accommodation account, deaccenting nonidentical material is ungrammatical, so there is a clearly defined possibility of an acceptability gap between these cases and canonical deaccenting under identity. The main puzzle is why deaccenting of inferable material is not “good” out of the blue, since accommodation accounts like Fox’s (2000) treat deaccented nonidentical constituents as accommodation-seeking material that should straightforwardly trigger construction of an alternative antecedent when the constituent’s meaning is lexically inferable from the antecedent. However, the experimental results in principle are highly compatible with this model. While it appears not to be correct that a deaccented inferable constituent itself can trigger accommodation of an alternative antecedent, at least in the cases investigated in the experiments here, the theory does allow for accommodation to occur more readily when the alternative antecedent has additional support in the discourse. This could explain the observed cline of acceptability for deaccented inferable material, with a supportive context and especially presupposition triggers helping listeners realize that inferable material was being felicitously marked as given in the discourse.

By contrast, it is less apparent how to reconcile the grammatical account with the observed data, although it is not impossible to do so. The grammatical account treats deaccenting as licensed, and even mandatory, for any constituent that is inferable from an antecedent.

As such, it does not predict that there should be a wide gap in the acceptability of sentences with canonical deaccenting under identity versus deaccenting of inferable material. In other words, deaccenting should be automatic in cases where a verb is lexically inferable from an antecedent, but in reality, participants required a high degree of additional contextual support to mark such deaccenting as acceptable.

The most apparent way to reconcile the grammatical account with the current data seems to be to posit that speakers and listeners were simply unaware of the critical lexical inferencing relations in most of the experiments. That is, if the grammatical deaccenting mechanism makes reference to inferencing relations, but participants did not calculate these relations while performing the task, then it is not surprising that they treated inferable verbs as though they were discourse-new. The presence of a supportive context or a presupposition trigger in Chapter 7 forced participants to calculate the inferencing relations in ways that were not necessary in the prior studies, causing a corresponding increase in the acceptability of deaccented inferable material. While this proposal seems on the surface to be at odds with the finding from Experiment 8 that all of the target inferable verbs were made highly available by their antecedents, it is possible that there is not a straightforward mapping between the ratings elicited in that experiment and the degree to which speakers and hearers are aware of such relations when dealing with prosody encoding in out-of-the-blue contexts.

8.3 Synthesis, discussion, and future directions

To close, it is worth considering the experimental results for verb phrase ellipsis and deaccenting together. Ellipsis and deaccenting have previously been treated together in the literature, so it is worth asking to what extent they can be given a uniform analysis in light of the new data presented here. In addition, consideration of multiple types of anaphoric or otherwise context-dependent constructions may yield new insights about the nature of such phenomena more generally.

For both VPE and deaccenting, there were well established questions in the prior liter-

ature as to the relative contributions of local linguistic antecedents and broader context to licensing and interpretation. For both phenomena, it has been proposed that exact formal identity with a linguistic antecedent plays a fundamental role – in the syntactic identity accounts for VPE, and the accommodation account of deaccenting, which posits that only deaccenting of antecedent-identical constituents is grammatical. On the other hand, both phenomena have also received analyses that allow for direct, holistic consideration of the discourse context at large – semantic identity or discourse-driven accounts for VPE, and the grammatical account of deaccenting licensing under nonidentity.

The results of the experiments presented here indicate that both verb phrase ellipsis and deaccenting involve substantial interaction between linguistic antecedents and broader discourse contexts. For VPE, interpretation was largely guided by the content of the linguistic antecedent, but could be affected by the broader context when the latter included salient information that was not available in the linguistic antecedent. In the case of deaccenting, inferable constituents were generally not deaccented in production, nor were they marked acceptable in perception in out-of-the-blue contexts. However, the acceptability of deaccented inferable material in perception increased as a function of contextual support for identical readings of the antecedent and target.

The studies presented here make important contributions to the empirical landscape for both VPE and deaccenting. For VPE, the studies addressed a type of discourse context that has received little attention in the literature, in which there is both a viable antecedent VP as well as salient information in the broader context. The finding that listeners consider all available information during interpretation is not surprising from a language comprehension point of view, and certainly does not invalidate prior work on narrower identity mechanisms, which largely considered exclusively linguistic discourse contexts. Rather, the research presented here constitutes a first exploration of interpretation in more complex discourse contexts, and the experimental results must be accounted for by any model of context-situated VPE interpretation that purports to be general to the type of naturalistic

discourse contexts that speakers and hearers encounter in everyday situations.

For deaccenting, the empirical findings interface more closely with judgments that were previously reported in the literature. Prior work has commonly cited deaccented inferable material as acceptable, but this was based on introspective judgments for a small number of examples and had not received rigorous empirical investigation. The deaccenting studies presented here clarified the empirical situation; deaccenting of inferable material can be felicitous, but it requires more support from the broader discourse context than has been appealed to in the literature. Further work may explain the discrepancy between the experimental results reported here and the judgments in the literature. For instance, prior authors may have implicitly considered additional discourse support for deaccenting that they did not include in their examples (e.g., examples of deaccenting under nonidentity are often accompanied by a text explanation of why the deaccenting “works”), or the experimental stimuli may simply be different in kind from prior examples in some unanticipated way. Nevertheless, the current work represents a large step forward in understanding the dynamics of deaccenting licensing in more naturalistic contexts.

One major question that remains is the extent to which antecedent/context interactions for VPE and deaccenting can be given a unified analysis, and in turn, whether this analysis can be unified with those for other anaphoric and context-sensitive constructions. In light of the experimental results, two possibilities remain open for a unified treatment of context for VPE and deaccenting.

First, it was noted that both phenomena are amenable to an analysis which privileges linguistic antecedents in the grammar. VPE interpretation was highly constrained by the form of the linguistic antecedent, so it may be best to suggest that interpretation proceeds according to strict identity with the antecedent, with a secondary mechanism that considers information from the broader context. The main challenge for this approach is to spell out this secondary mechanism, since according to the current data, it does not require a formal antecedent-ellipsis mismatch like other accommodation or repair mechanisms

in the literature. The linguistic antecedent also appears to be privileged in deaccenting. Antecedent-identical material is readily deaccented in production and is always marked as felicitous in production, even in out-of-the-blue contexts. By contrast, deaccenting inferable material requires considerable support from the discourse context. Thus, it may be best to propose that the grammar of deaccenting requires linguistic identity, and that deaccenting under nonidentity is due to a secondary operation like antecedent accommodation.

On the other hand, it was also noted that both phenomena could potentially be analyzed with a grammatical mechanism that considers the discourse more holistically. In the case of VPE, this would take the form of a semantic identity or otherwise discourse-driven account. Such models allow interpreters to consider both the content of the linguistic antecedent and the information in the broader discourse in interpretation, without drawing a principled distinction on the basis of the information's source. The major empirical challenge here is that the heuristics for identifying the locus of resolution are not clearly spelled out. The only proposal made so far that speaks to the distinction between a viable VP antecedent and broader contextual information holds that VP antecedents are more salient or cognitively available. However, the experimental results suggest that the linguistic antecedent constrains interpretation even beyond its contribution to salience, making additional work necessary to identify and test the relevant heuristics.

A parallel analysis of deaccenting licensing holds that the grammar does not require strict linguistic identity between antecedents and deaccenting targets, but rather that any material that is inferable from an antecedent, and possibly from the discourse at large, is a felicitous target for deaccenting. There are challenges for reconciling this account with the current data, but one reasonable proposal is that establishing inference relations is itself challenging or costly in some way, preventing apparent “automatic” licensing of deaccenting even in cases where it is grammatical.

Drawing a firm conclusion about these mechanisms, and the extent to which they are uniform for the two constructions, may inform the broader theory of anaphora and context

sensitivity. A widely cited proposal is that such constructions can be *surface anaphors*, meaning they depend grammatically on a linguistic antecedent for interpretation, or *deep anaphors*, meaning they are free to receive an interpretation from the information available in the discourse at large (Hankamer and Sag, 1976). A determination that the grammars of VPE and deaccenting require strict identity with an antecedent, with broader contextual information affecting interpretation only through an extragrammatical process, would support the notion that there is a meaningful divide between such surface anaphors on the one hand and deep anaphors like *do that*, *do it*, and pronouns on the other. A conclusion that the grammatical mechanisms of VPE and deaccenting actually consider all discourse information holistically would suggest that the deep-surface distinction is merely descriptive, with relatively “surface” anaphors introducing biases that prefer resolution with respect to antecedent information more strongly than relatively “deep” anaphors.

Future work can build on the present results in a number of ways. One attractive direction for further research is to investigate VPE and deaccenting interpretation with online methodologies that are sensitive to the time course of interpretation. For instance, a well-designed visual world eye-tracking methodology might be able to detect if antecedent-supported VPE interpretations are accessed prior to interpretations supported by the context at large. For deaccenting, a self-paced listening methodology (Ferreira et al., 1996a,b; Marinis, 2010) could potentially indicate whether processing deaccented inferable material is costlier than processing canonically deaccented identical material. In each case, a delay in accessing interpretations that require reference to the broad context might be used to argue that such information is considered only via a second-stage operation like accommodation. However, it may also be the case that such information is simply costly or otherwise time-intensive to access, with processing delays not necessarily indicating an extragrammatical process.

Another important direction for future research is to expand the range of constructions under investigation. The current dissertation briefly considered the verbal anaphor *do that* and compared its interpretation to that of VPE. Future work on this and other verbal

anaphors (e.g., *do it*) might shed light on more general interpretation processes and speak to the question of where on the deep-surface spectrum each construction lies, and why. For deaccenting, only very simple constituents have received thorough empirical investigation – verbs in the current dissertation, and nouns in Chodroff and Cole’s (2019) study. There may be further insights to be gained from examining the deaccenting of larger constituents. In addition, further work on deaccenting under co-reference, as well as deaccenting licensed by more informal world knowledge-driven inferences or by information from the nonlinguistic context, may shed additional light on the division of labor between grammar and extragrammatical processes in accent assignment.

Finally, it is worth considering how to investigate VPE and deaccenting in even more articulated discourses than were considered here. Prior studies have examined the interpretation of VPE in more complex linguistic contexts, such as cases with multiple viable VP antecedents (Hardt, 1993). Combining complex linguistic discourses with informative broader contexts may yield additional insight on the heuristics that guide recruitment of information from various sources, and may constitute a step toward more faithfully representing the complex discourses speakers and hearers are likely to encounter in day-to-day conversation. On the prosodic side, the current dissertation focused on close lexical inferring relations, but the literature also posits examples with more distant relations that rely on non-lexical world knowledge, such as the *call a Republican/insult* example (Lakoff, 1972) and even cases where a constituent is deaccented because its meaning is made available by the nonlinguistic context (Rochemont, 1986), and these examples are worthy of future systematic consideration. It is also puzzling that the introduction of *too* raised scores to a high level even in conditions with canonically “unlicensed” prosody, such as accented repeated material, and future research investigating the connection between presupposition triggers and prosody is warranted. Examining cases like those outlined here may shed further light on the interactions between overt antecedents and the broader context in determining the felicitousness of various prosodic encodings.

As this dissertation has shown, verb phrase ellipsis and deaccenting both instantiate complex interactions between local linguistic antecedents and broader discourse contexts. The current work has shed considerable light on the empirical landscape of this interaction for both constructions and helped constrain the space of hypothetical mechanisms underlying these interactions. At the same time, many questions remain open pertaining to both the empirical facts of interpretation in context and the best theoretical architectures to capture the data. There is substantial work left to do to draw firm conclusions about the nature of the licensing and interpretation of VPE and deaccenting, the extent to which their grammatical mechanisms are related to one another, and their connections to the broader category of anaphoric and context-sensitive expressions. Future experimental and theoretical work in this domain will no doubt be fruitful.

APPENDIX A

EXPERIMENT 1: VPE INTERPRETATION IN CONTEXT

A.1 Task and layout

In an office



Standing worker: I'm going to take four candies!

Sitting worker: You can't.

On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the sitting worker meant:

You can't take any candies.

(Unlikely) (Likely)

Use number keys or click boxes to answer.

Figure A.1: Sample screen from Experiment 1

A.2 Participant statistics and data loss

Participants:	195
Gender:	113 female, 82 male
Mean age:	35.7 years
Non-native speakers excluded:	3 (1.5%)
Attention criterion:	Participant's mean reaction time ≤ 2000 ms
Inattentive participants excluded:	3 (1.5%)
Average completion time (included participants):	11 min, 28 sec
Payment:	USD 1.50

A.3 Critical items

For the Chapter 3 experiments, the Latin square functionality of Ibex Farm was implemented in such a way that several consecutive participants could be assigned to the same list. As a result, some individual item/condition cells have very few observations, including some with only one observation. This should be taken into account when considering possible by-item differences in responses.

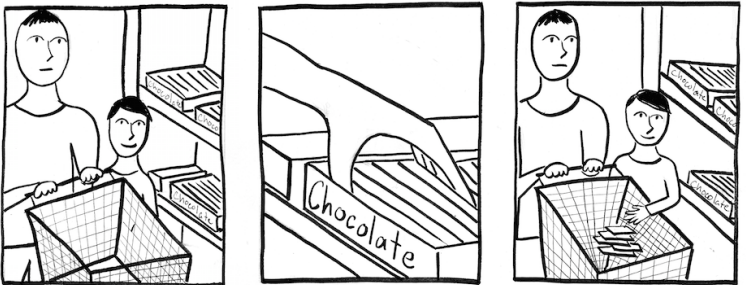
Item 1

Title At the grocery store

Unavailable comic



Available comic



Salient comic



Unmodified antecedent

Son: I want to buy candy bars!

Modified antecedent

Son: I want to buy five candy bars!

Reply

Father: We can't.

Polar prompt

On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the father meant: We can't buy any candy bars today.

Scalar prompt

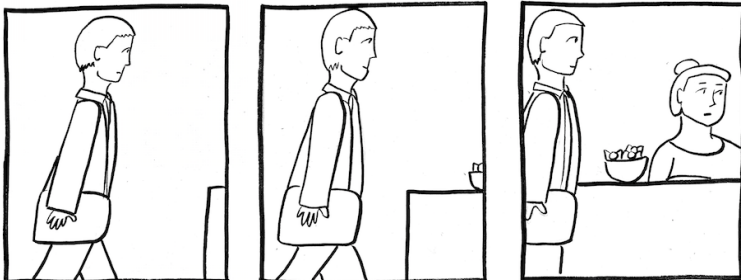
On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the father meant: We can't buy five candy bars today, but maybe we could get fewer.

Item 2

Title

In an office

Unavailable comic



Available comic



Salient comic



Unmodified antecedent
Modified antecedent
Reply

Polar prompt

Scalar prompt

Standing worker: I'm going to take candies!
Standing worker: I'm going to take four candies!
Sitting worker: You can't.
On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the sitting worker meant: You can't take any candies.
On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the sitting worker meant: We can't buy four candies, but maybe you can take fewer.

Item 3

Title

In a cooking class

Unavailable comic



Available comic



Salient comic



Unmodified antecedent
Modified antecedent
Reply

Polar prompt

Student: Now I'm going to add chili peppers.
Student: Now I'm going to add four chili peppers.
Instructor: You shouldn't.
On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the instructor meant: You shouldn't add any chili peppers.

Scalar prompt

On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the instructor meant: You shouldn't add four chili peppers, but maybe fewer would be OK.

Item 4

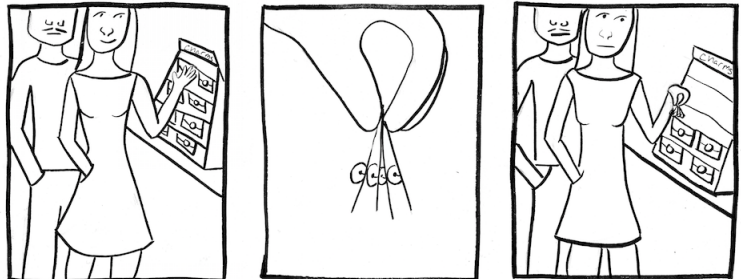
Title

In a jewelry store

Unavailable comic



Available comic



Salient comic



Unmodified antecedent

Wife: I want to get charms for my bracelet.

Modified antecedent

Wife: I want to get four charms for my bracelet.

Reply

Husband: You shouldn't.

Polar prompt

On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the husband meant: You shouldn't buy any bracelet charms today.

Scalar prompt

On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the husband meant: You shouldn't buy four bracelet charms today, but maybe fewer would be OK.

Item 5

Title

In a clothing store

Unavailable comic



Available comic



Salient comic



Unmodified antecedent

Salesperson: I think you should buy shirts today.

Modified antecedent

Salesperson: I think you should buy three shirts today.

Reply

Customer: I shouldn't.

Polar prompt

On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the customer meant: I shouldn't buy any shirts today.

Scalar prompt

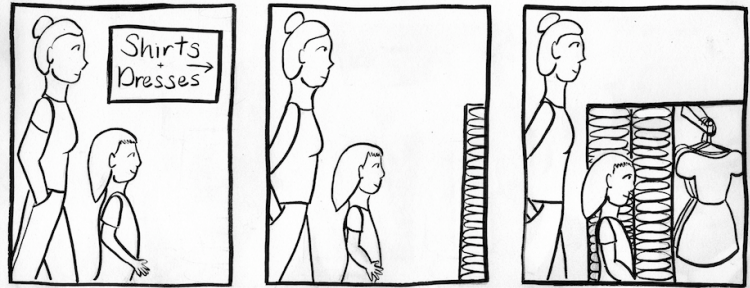
On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the father meant: I shouldn't buy three shirts today, but maybe fewer would be OK.

Item 6

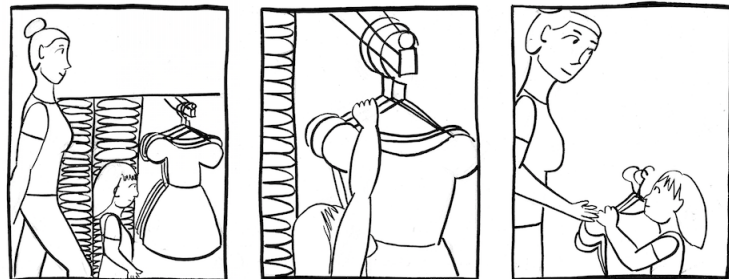
Title

In a children's store

Unavailable comic



Available comic



Salient comic



Unmodified antecedent

Daughter: I want to buy princess dresses!

Modified antecedent

Daughter: I want to buy three princess dresses!

Reply

Mother: We can't.

Polar prompt

On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the mother meant: We can't buy any princess dresses today.

Scalar prompt

On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the mother meant: We can't buy three princess dresses today, but maybe we could get fewer.

A.4 By-item results

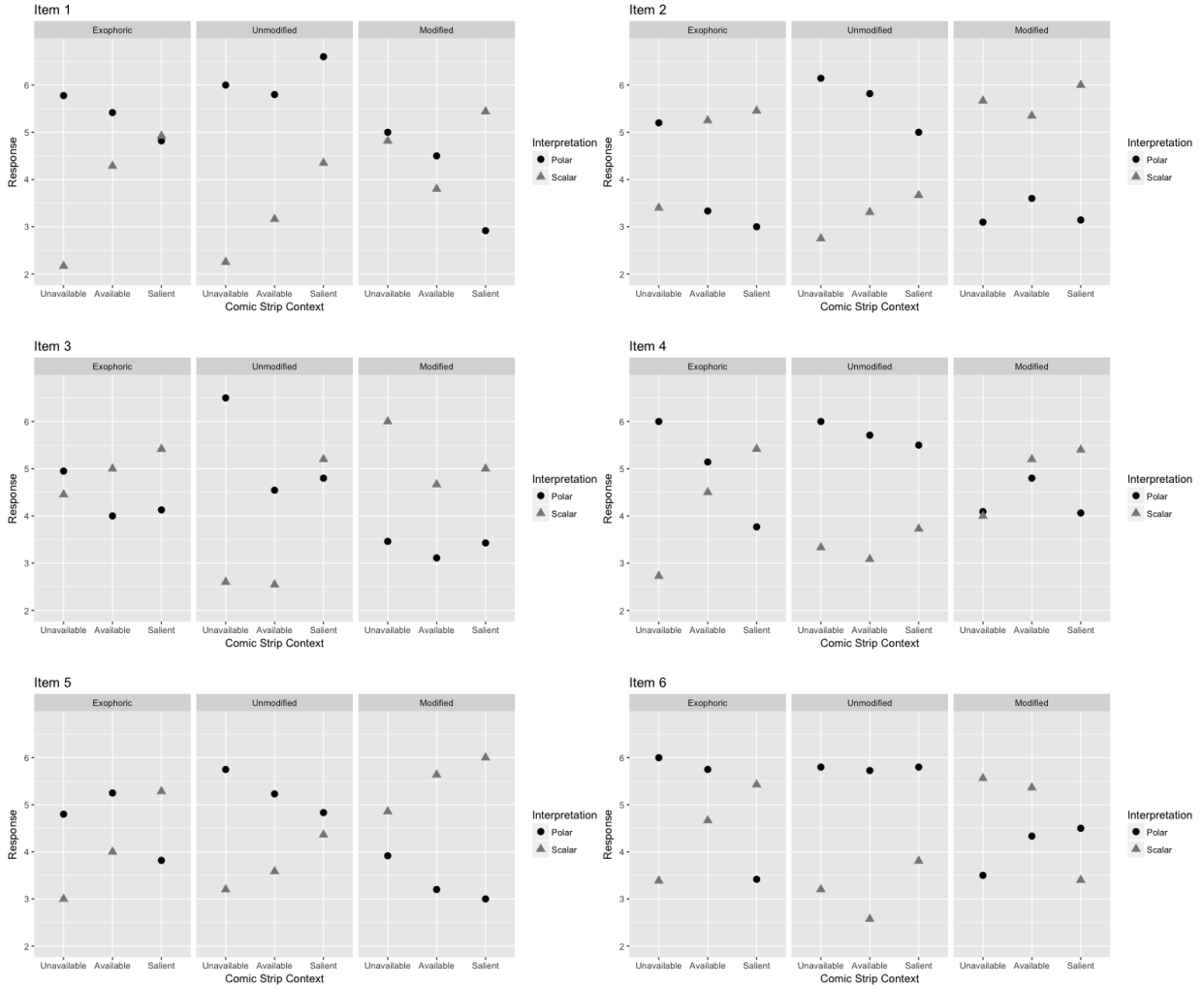


Figure A.2: Experiment 1 results by item

A.5 Fillers

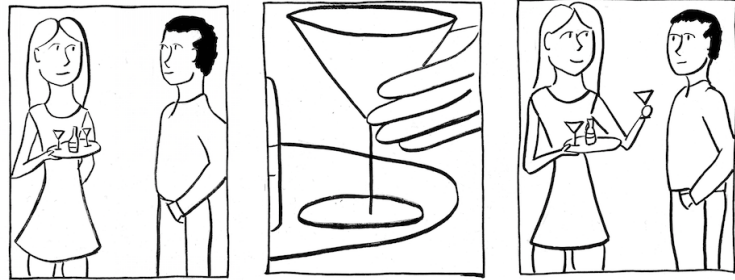
Each participant saw each filler item once. For each participant, 5 of the filler items used the polar prompt and 5 used the scalar prompt. The results were not analyzed. The mismatch between the comic (writing with pencils) and the utterances (handing a hammer) in Item 16 was due to error during experiment construction.

Item 7

Title

At a party

Comic



Antecedent

Host: Do you want another drink?

Reply

Guest: I really can't.

Polar prompt

On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the guest meant: I can't have another drink of any kind.

Scalar prompt

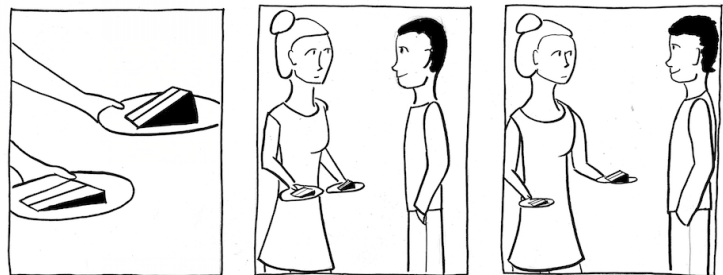
On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the guest meant: I can't have another cocktail, but maybe a beer would be OK.

Item 8

Title

At a party

Comic



Antecedent

Host: Do you want a big piece of cake?

Reply

Guest: I shouldn't.

Polar prompt

On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the guest meant: I shouldn't have any cake at all.

Scalar prompt

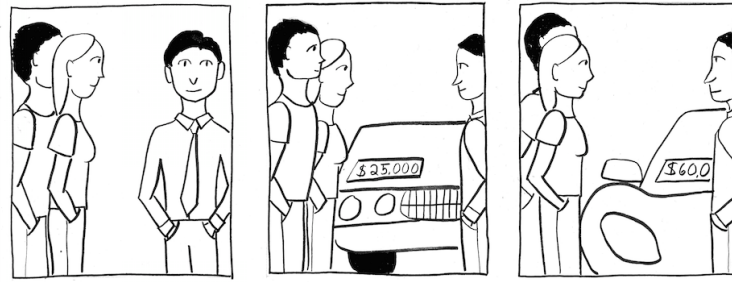
On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the guest meant: I shouldn't have a big piece of cake, but maybe a smaller piece would be OK.

Item 9

Title

At a car dealership

Comic



Antecedent

Salesperson: I see you're looking at the fast car.

Reply

Wife: We can't.

Polar prompt

On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the wife meant: We can't buy a car at all today.

Scalar prompt

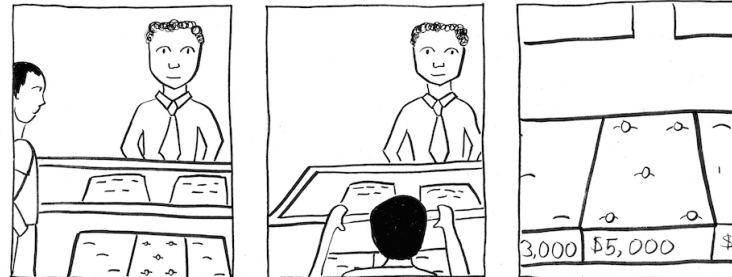
On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the wife meant: We can't buy a fast car today, but we might be willing to buy a slower car.

Item 10

Title

In a jewelry store

Comic



Antecedent

Salesperson: Were you looking to buy a ring today?

Reply

Customer: I can't.

Polar prompt

On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the customer meant: I can't buy a ring at all today.

Scalar prompt

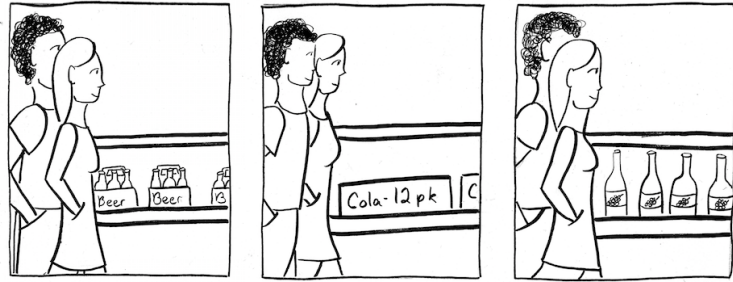
On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the customer meant: I can't buy an expensive ring today, but a cheaper one might be OK.

Item 11

Title

In a liquor store

Comic



Antecedent

Man: Should we buy the expensive wine?

Reply

Woman: We shouldn't.

Polar prompt

On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the woman meant: We shouldn't buy any wine today.

Scalar prompt

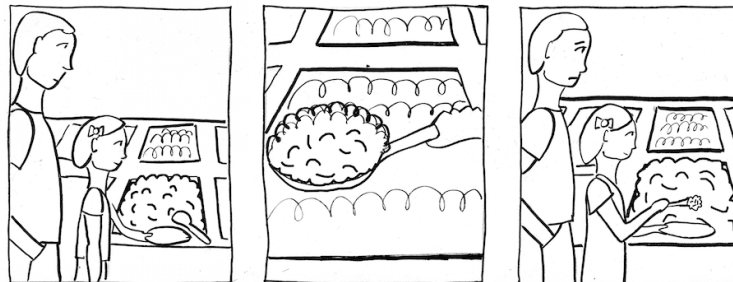
On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the woman meant: We shouldn't buy the expensive wine today, but maybe a cheaper wine would be OK.

Item 12

Title

At a buffet

Comic



Statement

Father: Don't!

Polar prompt

On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the father meant: Don't take any of that.

Scalar prompt

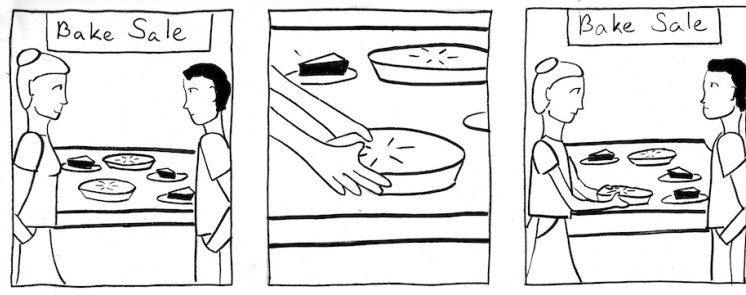
On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the father meant: Don't take so much of that.

Item 13

Title

At a bake sale

Comic



Antecedent

Woman: I'm going to buy a whole pie!

Reply

Man: Don't!

Polar prompt

On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the man meant: Don't buy any pie at all.

Scalar prompt

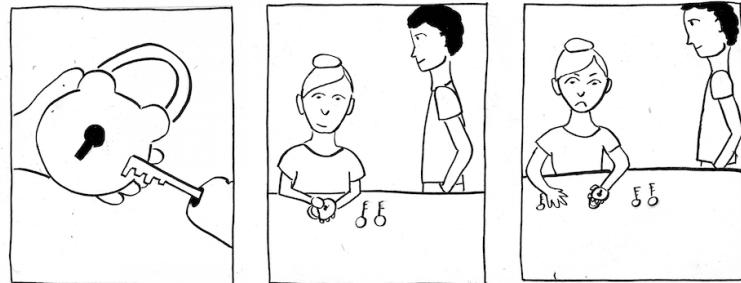
On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the man meant: Don't buy a whole pie, but maybe a slice of pie would be OK.

Item 14

Title

Opening a lock

Comic



Statement

Standing person: You can't.

Polar prompt

On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the standing person meant: You can't open that lock at all.

Scalar prompt

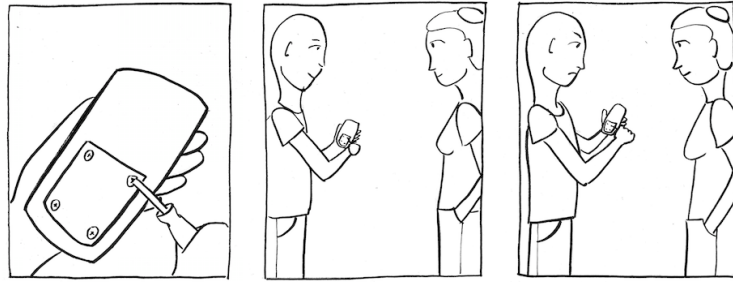
On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the standing person meant: You can't open that lock with that key.

Item 15

Title

Turning a screw

Comic



Statement

Standing person: You can't.

Polar prompt

On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the standing person meant: You can't turn that screw at all.

Scalar prompt

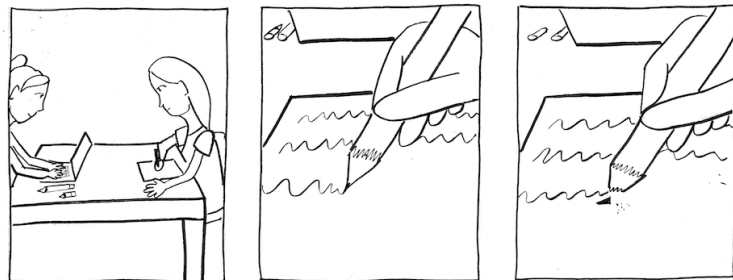
On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the standing person meant: You can't turn that screw with that screwdriver.

Item 16

Title

Working with tools

Comic



Antecedent

Person on right: Can you hand me the big hammer?

Reply

Person on left: I can't.

Polar prompt

On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the person on the left meant: I can't hand you any hammer at all.

Scalar prompt

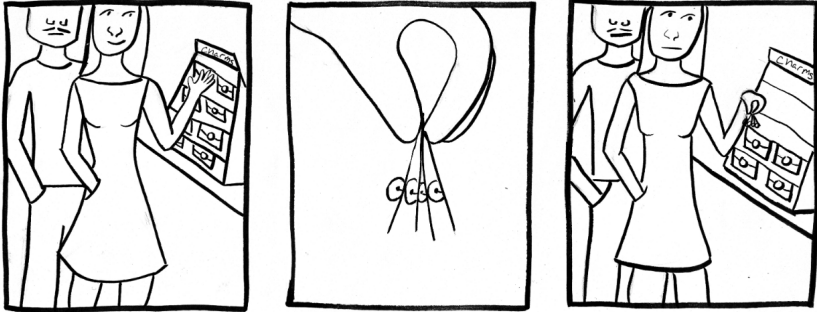
On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the person on the left meant: I can't hand you the big hammer, but I can hand you a smaller hammer.

APPENDIX B

EXPERIMENT 2: ELLIPSIS-SPECIFIC INTERPRETATION EFFECTS

B.1 Task and layout

In a jewelry store



Wife: I want to get charms for my bracelet.

Husband: You shouldn't get charms for your bracelet.

On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the husband meant:

You shouldn't buy any bracelet charms today.

(Unlikely) (Likely)

Use number keys or click boxes to answer.

Figure B.1: Sample screen from Experiment 2

B.2 Participant statistics and data loss

Participants:	204
Gender:	110 female, 94 male
Mean age:	34.3 years
Non-native speakers excluded:	6 (2.9%)
Attention criterion:	Participant's mean reaction time ≤ 2000 ms
Inattentive participants excluded:	3 (1.5%)
Average completion time (included participants):	9 min, 32 sec
Payment:	USD 1.50

B.3 Critical items

For the Chapter 3 experiments, the Latin square functionality of Ibex Farm was implemented in such a way that several consecutive participants could be assigned to the same list. As a result, some individual item/condition cells have very few observations, including some with only one observation. This should be taken into account when considering possible by-item differences in responses. The comic strips were identical to Experiment 1. See Appendix A for comics.

Item 1

Title	At the grocery store
Exophoric reply	<i>Father:</i> We can't.
Unmodified antecedent/reply	<i>Son:</i> I want to buy candy bars! <i>Father:</i> We can't buy candy bars.
Modified antecedent/reply	<i>Son:</i> I want to buy five candy bars! <i>Father:</i> We can't buy five candy bars.
Polar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the father meant: We can't buy any candy bars today.
Scalar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the father meant: We can't buy five candy bars today, but maybe we could get fewer.

Item 2

Title	In an office
Exophoric reply	<i>Sitting worker:</i> You can't.

Unmodified antecedent/reply	<i>Standing worker:</i> I'm going to take candies! <i>Sitting worker:</i> You can't take candies.
Modified antecedent/reply	<i>Standing worker:</i> I'm going to take four candies! <i>Sitting worker:</i> You can't take four candies.
Polar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the sitting worker meant: You can't take any candies.
Scalar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the sitting worker meant: We can't buy four candies, but maybe you can take fewer.

Item 3

Title	In a cooking class
Exophoric reply	<i>Instructor:</i> You shouldn't.
Unmodified antecedent/reply	<i>Student:</i> Now I'm going to add chili peppers. <i>Instructor:</i> You shouldn't add chili peppers.
Modified antecedent/reply	<i>Student:</i> Now I'm going to add four chili peppers. <i>Instructor:</i> You shouldn't add four chili peppers.
Polar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the instructor meant: You shouldn't add any chili peppers.
Scalar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the instructor meant: You shouldn't add four chili peppers, but maybe fewer would be OK.

Item 4

Title	In a jewelry store
Exophoric reply	<i>Husband:</i> You shouldn't.
Unmodified antecedent/reply	<i>Wife:</i> I want to get charms for my bracelet. <i>Husband:</i> You shouldn't get charms for your bracelet.
Modified antecedent/reply	<i>Wife:</i> I want to get four charms for my bracelet. <i>Husband:</i> You shouldn't get four charms for your bracelet.
Polar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the husband meant: You shouldn't buy any bracelet charms today.
Scalar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the husband meant: You shouldn't buy four bracelet charms today, but maybe fewer would be OK.

Item 5

Title	In a clothing store
Exophoric reply	<i>Customer:</i> I shouldn't.
Unmodified antecedent/reply	<i>Salesperson:</i> I think you should buy shirts today. <i>Customer:</i> I shouldn't buy shirts today.
Modified antecedent/reply	<i>Salesperson:</i> I think you should buy three shirts today. <i>Customer:</i> I shouldn't buy three shirts today.
Polar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the customer meant: I shouldn't buy any shirts today.
Scalar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the father meant: I shouldn't buy three shirts today, but maybe fewer would be OK.

Item 6

Title	In a children's store
Exophoric reply	<i>Mother:</i> We can't.
Unmodified antecedent/reply	<i>Daughter:</i> I want to buy princess dresses! <i>Mother:</i> We can't buy princess dresses.
Modified antecedent/reply	<i>Daughter:</i> I want to buy three princess dresses! <i>Mother:</i> We can't buy three princess dresses.
Polar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the mother meant: We can't buy any princess dresses today.
Scalar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the mother meant: We can't buy three princess dresses today, but maybe we could get fewer.

B.4 By-item results

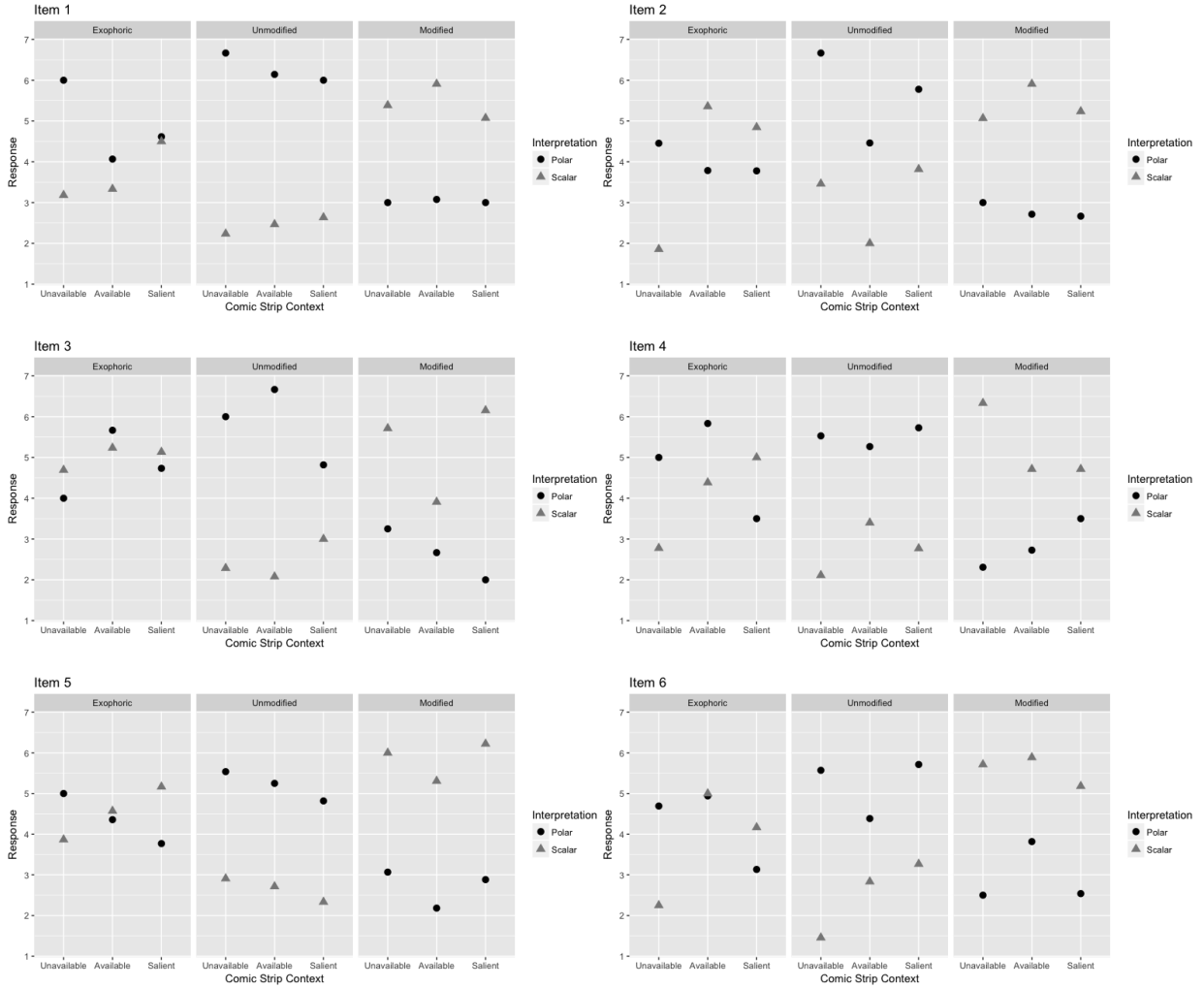


Figure B.2: Experiment 2 results by item

B.5 Fillers

Each participant saw each filler item once. For each participant, 5 of the filler items used the polar prompt and 5 used the scalar prompt. The results were not analyzed, but are available in the accompanying results file. The comic strips were identical to Experiment 1. See Appendix A for comics.

Item 7

Title	At a party
Antecedent	<i>Host:</i> Do you want another drink?
Reply	<i>Guest:</i> I really can't have another drink.
Polar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the guest meant: I can't have another drink of any kind.
Scalar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the guest meant: I can't have another cocktail, but maybe a beer would be OK.

Item 8

Title	At a party
Antecedent	<i>Host:</i> Do you want a big piece of cake?
Reply	<i>Guest:</i> I shouldn't have a big piece of cake.
Polar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the guest meant: I shouldn't have any cake at all.
Scalar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the guest meant: I shouldn't have a big piece of cake, but maybe a smaller piece would be OK.

Item 9

Title	At a car dealership
Antecedent	<i>Salesperson:</i> I see you're looking at the fast car.
Reply	<i>Wife:</i> We can't look at the fast car.
Polar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the wife meant: We can't buy a car at all today.
Scalar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the wife meant: We can't buy a fast car today, but we might be willing to buy a slower car.

Item 10

Title	In a jewelry store
Antecedent	<i>Salesperson:</i> Were you looking to buy a ring today?
Reply	<i>Customer:</i> I can't buy a ring today.
Polar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the customer meant: I can't buy a ring at all today.

Scalar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the customer meant: I can't buy an expensive ring today, but a cheaper one might be OK.
Item 11	
Title	In a liquor store
Antecedent	<i>Man:</i> Should we buy the expensive wine?
Reply	<i>Woman:</i> We shouldn't buy the expensive wine.
Polar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the woman meant: We shouldn't buy any wine today.
Scalar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the woman meant: We shouldn't buy the expensive wine today, but maybe a cheaper wine would be OK.
Item 12	
Title	At a buffet
Statement	<i>Father:</i> Don't!
Polar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the father meant: Don't take any of that.
Scalar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the father meant: Don't take so much of that.
Item 13	
Title	At a bake sale
Antecedent	<i>Woman:</i> I'm going to buy a whole pie!
Reply	<i>Man:</i> Don't buy a whole pie!
Polar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the man meant: Don't buy any pie at all.
Scalar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the man meant: Don't buy a whole pie, but maybe a slice of pie would be OK.
Item 14	
Title	Opening a lock
Statement	<i>Standing person:</i> You can't.
Polar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the standing person meant: You can't open that lock at all.

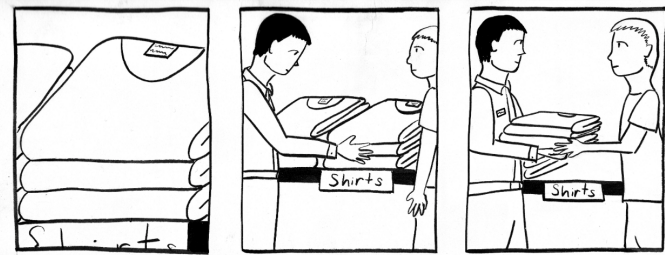
Scalar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the standing person meant: You can't open that lock with that key.
Item 15	
Title	Turning a screw
Statement	<i>Standing person:</i> You can't.
Polar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the standing person meant: You can't turn that screw at all.
Scalar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the standing person meant: You can't turn that screw with that screwdriver.
Item 16	
Title	Writing
Antecedent	<i>Person on right:</i> Can you hand me the short pencil?
Reply	<i>Person on left:</i> I can't hand you the short pencil.
Polar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the person on the left meant: I can't hand you any pencil at all.
Scalar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the person on the left meant: I can't hand you the short pencil, but maybe I can hand you a longer pencil.

APPENDIX C

EXPERIMENT 3: DISCOURSE COHERENCE OF VPE INTERPRETATIONS

C.1 Task and layout

In a clothing store



Salesperson: I think you should buy three shirts today.

Customer: I shouldn't buy three shirts today, but maybe fewer would be OK.

On a scale of 1 to 7, where 1 is the least coherent and 7 is the most coherent, **how coherent** is the customer's statement considering what has come before?

(Incoherent) (Coherent)

Use number keys or click boxes to answer.

Figure C.1: Sample screen from Experiment 3

C.2 Participant statistics and data loss

Participants:	196
Gender:	108 female, 88 male
Mean age:	35.1 years
Non-native speakers excluded:	4 (2.0%)
Attention criterion:	Participant's mean reaction time $\leq$ 2000 ms
Additional inattentive participants excluded:	6 (3.1%)
Average completion time (included participants):	10 min, 36 sec
Payment:	USD 1.50
	400

C.3 Critical items

For the Chapter 3 experiments, the Latin square functionality of Ibex Farm was implemented in such a way that several consecutive participants could be assigned to the same list. As a result, some individual item/condition cells have very few observations, including some with only one observation. This should be taken into account when considering possible by-item differences in responses. The comic strip contexts are identical to those used in Experiment 1. See Appendix A for comics.

Item 1

Title	At the grocery store
Unmodified antecedent	<i>Son:</i> I want to buy candy bars!
Modified antecedent	<i>Son:</i> I want to buy five candy bars!
Polar reply	<i>Father:</i> We can't buy any candy bars today.
Scalar reply	<i>Father:</i> We can't buy five candy bars today, but maybe we could get fewer.
Prompt	On a scale from 1 to 7, where 1 is the least coherent and 7 is the most coherent, how coherent is the father's statement considering what came before?

Item 2

Title	In an office
Unmodified antecedent	<i>Standing worker:</i> I'm going to take candies!
Modified antecedent	<i>Standing worker:</i> I'm going to take four candies!
Reply	<i>Sitting worker:</i> You can't take any candies.
Reply	<i>Sitting worker:</i> You can't take four candies, but maybe you can take fewer.
Prompt	On a scale from 1 to 7, where 1 is the least coherent and 7 is the most coherent, how coherent is the sitting worker's statement considering what came before?

Item 3

Title	In a cooking class
Unmodified antecedent	<i>Student:</i> Now I'm going to add chili peppers.
Modified antecedent	<i>Student:</i> Now I'm going to add four chili peppers.
Reply	<i>Instructor:</i> You shouldn't add any chili peppers.
Reply	<i>Instructor:</i> You shouldn't add four chili peppers, but maybe fewer would be OK.
Prompt	On a scale from 1 to 7, where 1 is the least coherent and 7 is the most coherent, how coherent is the instructor's statement considering what came before?

Item 4

Title	In a jewelry store
Unmodified antecedent	<i>Wife:</i> I want to get charms for my bracelet.
Modified antecedent	<i>Wife:</i> I want to get four charms for my bracelet.
Reply	<i>Husband:</i> You shouldn't buy any bracelet charms today.
Reply	<i>Husband:</i> You shouldn't buy four bracelet charms today, but maybe fewer would be OK.
Prompt	On a scale from 1 to 7, where 1 is the least coherent and 7 is the most coherent, how coherent is the husband's statement considering what came before?

Item 5

Title	In a clothing store
Unmodified antecedent	<i>Salesperson:</i> I think you should buy shirts today.
Modified antecedent	<i>Salesperson:</i> I think you should buy three shirts today.
Reply	<i>Customer:</i> I shouldn't buy any shirts today.
Reply	<i>Customer:</i> I shouldn't buy three shirts today, but maybe fewer would be OK.
Prompt	On a scale from 1 to 7, where 1 is the least coherent and 7 is the most coherent, how coherent is the customer's statement considering what came before?

Item 6

Title	In a children's store
Unmodified antecedent	<i>Daughter:</i> I want to buy princess dresses!
Modified antecedent	<i>Daughter:</i> I want to buy three princess dresses!
Reply	<i>Mother:</i> We can't buy any princess dresses today.
Reply	<i>Mother:</i> We can't buy three princess dresses today, but maybe we could get fewer.
Prompt	On a scale from 1 to 7, where 1 is the least coherent and 7 is the most coherent, how coherent is the mother's statement considering what came before?

C.4 By-item results

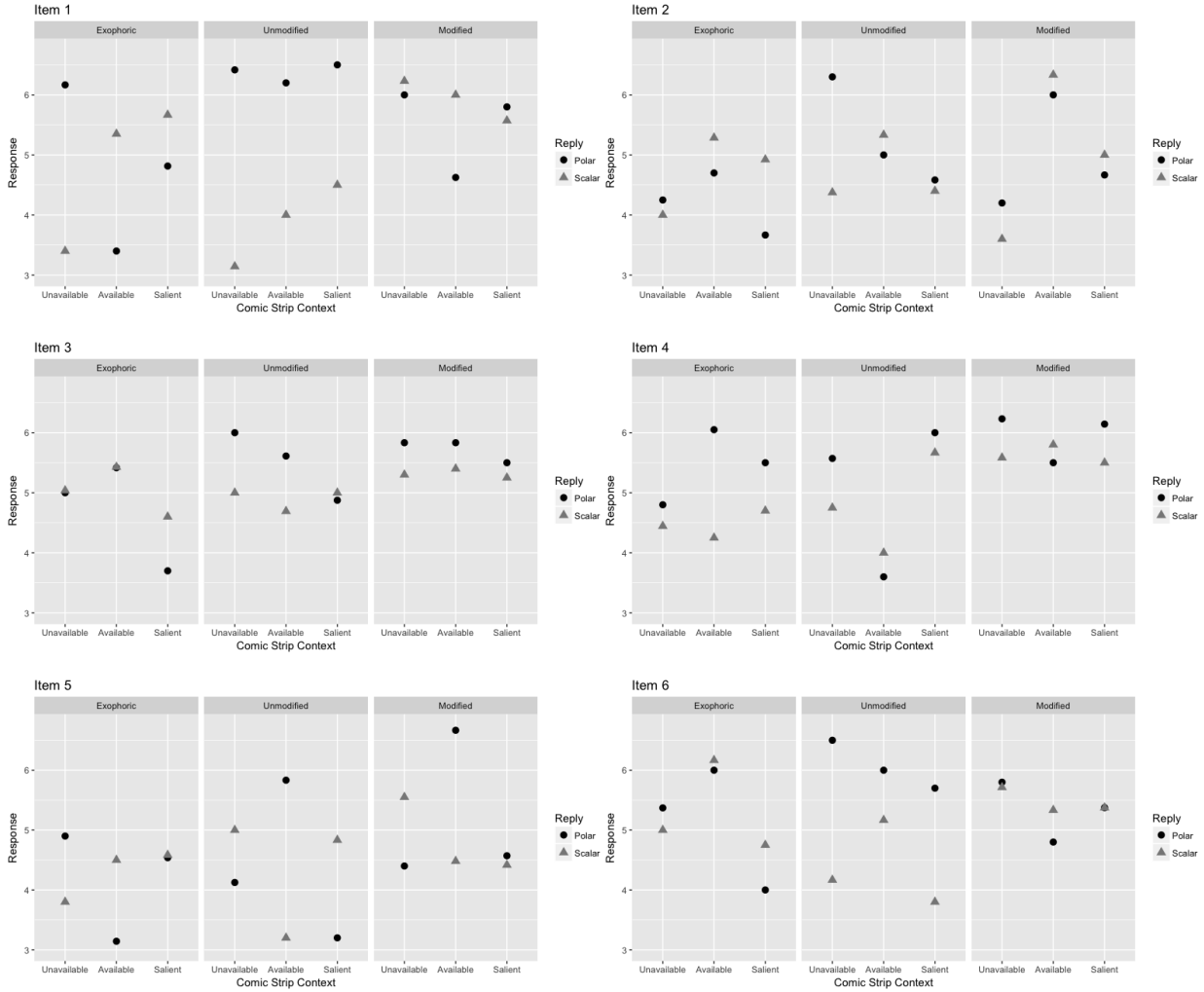


Figure C.2: Experiment 3 results by item

C.5 Fillers

Each participant saw each filler item once. For each participant, 5 of the filler items used the polar reply/statement and 5 used the scalar reply/statement. The results were not analyzed. The comic strip contexts are identical to those used in Experiment 1. See Appendix A for comics.

Item 7

Title At a party
Antecedent *Host:* Do you want another drink?
Polar reply *Guest:* I really can't have another drink of any kind.
Scalar reply *Guest:* I really can't have another cocktail, but maybe a beer would be OK.
Prompt On a scale from 1 to 7, where 1 is the least coherent and 7 is the most coherent, **how coherent** is the guest's statement considering what has come before?

Item 8

Title At a party
Antecedent *Host:* Do you want a big piece of cake?
Polar reply *Guest:* I shouldn't have any cake at all.
Scalar reply *Guest:* I shouldn't have a big piece of cake, but maybe a smaller piece would be OK.
Prompt On a scale from 1 to 7, where 1 is the least coherent and 7 is the most coherent, **how coherent** is the guest's statement considering what has come before?

Item 9

Title At a car dealership
Antecedent *Salesperson:* I see you're looking at the fast car.
Polar reply *Wife:* We can't buy a car at all today.
Scalar reply *Wife:* We can't buy a fast car today, but we might be willing to buy a slower car.
Prompt On a scale from 1 to 7, where 1 is the least coherent and 7 is the most coherent, **how coherent** is the wife's statement considering what has come before?

Item 10

Title In a jewelry store
Antecedent *Salesperson:* Were you looking to buy a ring today?
Polar reply *Customer:* I can't buy a ring at all today.
Scalar reply *Customer:* I can't buy an expensive ring today, but maybe I could buy a cheaper one.
Prompt On a scale from 1 to 7, where 1 is the least coherent and 7 is the most coherent, **how coherent** is the customer's statement considering what has come before?

Item 11

Title In a liquor store
Antecedent *Man:* Should we buy the expensive wine?
Polar reply *Woman:* We shouldn't buy any wine today.

Scalar reply *Woman:* We shouldn't buy the expensive wine today, but maybe a cheaper wine would be OK.

Prompt On a scale from 1 to 7, where 1 is the least coherent and 7 is the most coherent, **how coherent** is the woman's statement considering what has come before?

Item 12

Title At a buffet

Polar statement *Father:* Don't take any of that!

Scalar statement *Father:* Don't take so much of that!

Prompt On a scale from 1 to 7, where 1 is the least coherent and 7 is the most coherent, **how coherent** is the father's statement considering what has come before?

Item 13

Title At a bake sale

Antecedent *Woman:* I'm going to buy a whole pie!

Polar reply *Man:* Don't buy any pie at all.

Scalar reply *Man:* Don't buy a whole pie, but maybe a slice of pie would be OK.

Prompt On a scale from 1 to 7, where 1 is the least coherent and 7 is the most coherent, **how coherent** is the man's statement considering what has come before?

Item 14

Title Opening a lock

Polar statement *Standing person:* You can't open that lock at all.

Scalar statement *Standing person:* You can't open that lock with that key.

Prompt On a scale from 1 to 7, where 1 is the least coherent and 7 is the most coherent, **how coherent** is the standing person's statement considering what has come before?

Item 15

Title Turning a screw

Polar statement *Standing person:* You can't turn that screw at all.

Scalar statement *Standing person:* You can't turn that screw with that screwdriver.

Prompt On a scale from 1 to 7, where 1 is the least coherent and 7 is the most coherent, **how coherent** is the standing person's statement considering what has come before?

Item 16

Title Writing

Antecedent *Person on right:* Can you hand me the short pencil?

Polar reply

Scalar reply

Prompt

Person on left: I can't hand you any pencil at all.

Person on left: I can't hand you the short pencil, but maybe I can hand you a longer pencil.

On a scale from 1 to 7, where 1 is the least coherent and 7 is the most coherent, **how coherent** is the person on the left's statement considering what has come before?

APPENDIX D

EXPERIMENT 4: INTERPRETATION OF *DO THAT* IN CONTEXT

D.1 Task and layout

In a cooking class



Student: Now I'm going to add chili peppers.

Instructor: You shouldn't do that.

On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the instructor meant:

You shouldn't add any chili peppers.

(Unlikely) (Likely)

Use number keys or click boxes to answer.

Figure D.1: Sample screen from Experiment 4

D.2 Participant statistics and data loss

Participants:	102
Gender:	54 female, 48 male
Mean age:	30.4 years
Non-native speakers excluded:	5 (4.9%)
Attention criterion:	Participant’s mean reaction time ≤ 2000 ms
Inattentive participants excluded:	1 (1.0%)
Average completion time (included participants):	10 min, 19 sec
Payment:	USD 1.50

D.3 Critical items

For the Chapter 3 experiments, the Latin square functionality of Ibex Farm was implemented in such a way that several consecutive participants could be assigned to the same list. As a result, some individual item/condition cells have very few observations, including some with only one observation. This should be taken into account when considering possible by-item differences in responses. The comic strip contexts are identical to those used in Experiment 1. See Appendix A for comics.

Item 1

Title	At the grocery store
Unmodified antecedent	<i>Son</i> : I want to buy candy bars!
Modified antecedent	<i>Son</i> : I want to buy five candy bars!
Reply	<i>Father</i> : We can’t do that.
Polar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the father meant: We can’t buy any candy bars today.
Scalar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the father meant: We can’t buy five candy bars today, but maybe we could get fewer.

Item 2

Title	In an office
Unmodified antecedent	<i>Standing worker</i> : I’m going to take candies!
Modified antecedent	<i>Standing worker</i> : I’m going to take four candies!
Reply	<i>Sitting worker</i> : You can’t do that.

Polar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the sitting worker meant: You can't take any candies.
Scalar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the sitting worker meant: We can't buy four candies, but maybe you can take fewer.
Item 3	
Title	In a cooking class
Unmodified antecedent	<i>Student:</i> Now I'm going to add chili peppers.
Modified antecedent	<i>Student:</i> Now I'm going to add four chili peppers.
Reply	<i>Instructor:</i> You shouldn't do that.
Polar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the instructor meant: You shouldn't add any chili peppers.
Scalar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the instructor meant: You shouldn't add four chili peppers, but maybe fewer would be OK.
Item 4	
Title	In a jewelry store
Unmodified antecedent	<i>Wife:</i> I want to get charms for my bracelet.
Modified antecedent	<i>Wife:</i> I want to get four charms for my bracelet.
Reply	<i>Husband:</i> You shouldn't do that.
Polar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the husband meant: You shouldn't buy any bracelet charms today.
Scalar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the husband meant: You shouldn't buy four bracelet charms today, but maybe fewer would be OK.
Item 5	
Title	In a clothing store
Unmodified antecedent	<i>Salesperson:</i> I think you should buy shirts today.
Modified antecedent	<i>Salesperson:</i> I think you should buy three shirts today.
Reply	<i>Customer:</i> I shouldn't do that.
Polar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the customer meant: I shouldn't buy any shirts today.

Scalar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the father meant: I shouldn't buy three shirts today, but maybe fewer would be OK.
Item 6	
Title	In a children's store
Unmodified antecedent	<i>Daughter:</i> I want to buy princess dresses!
Modified antecedent	<i>Daughter:</i> I want to buy three princess dresses!
Reply	<i>Mother:</i> We can't do that.
Polar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the mother meant: We can't buy any princess dresses today.
Scalar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the mother meant: We can't buy three princess dresses today, but maybe we could get fewer.

D.4 By-item results

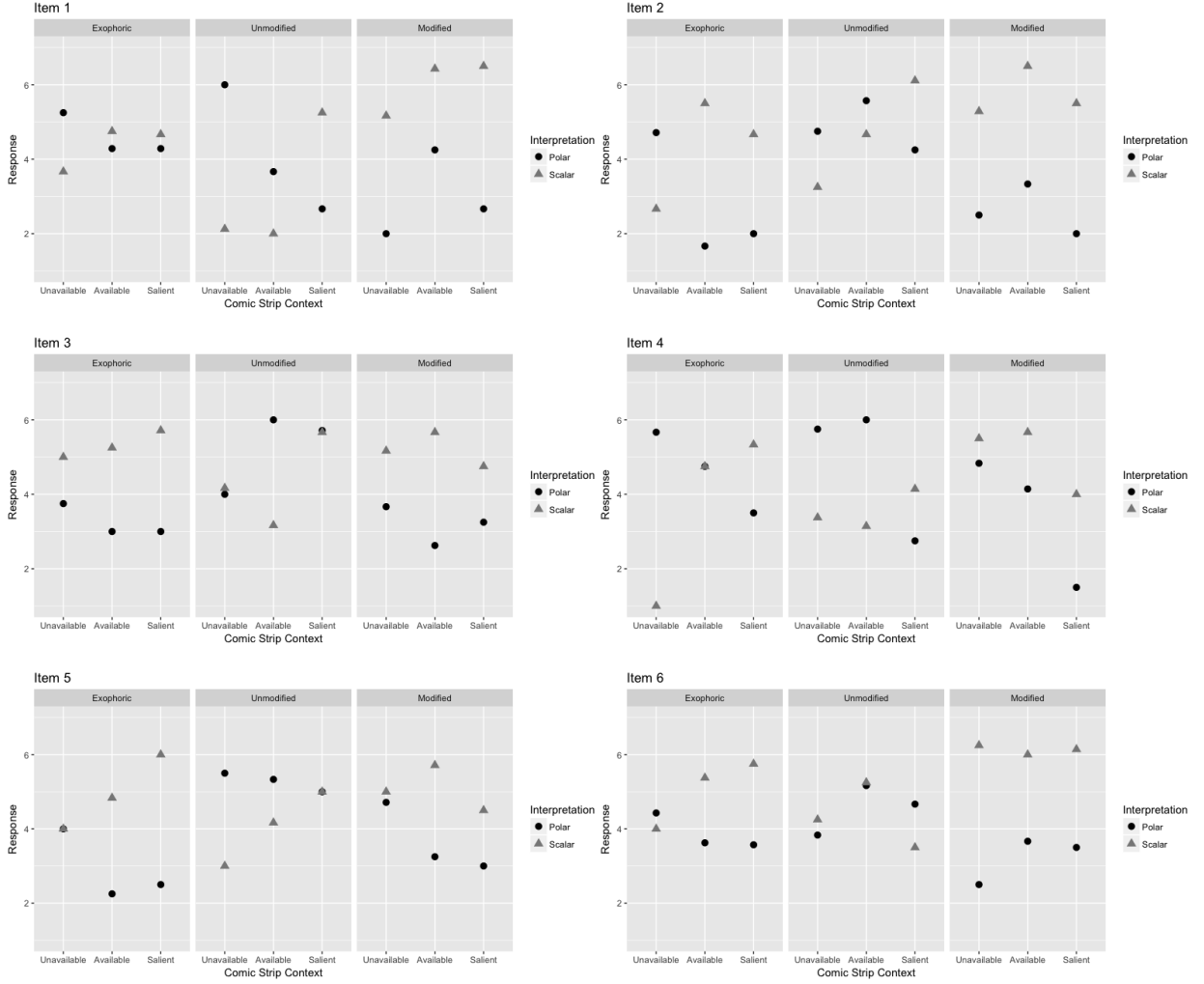


Figure D.2: Experiment 4 results by item

D.5 Fillers

Each participant saw each filler item once. For each participant, 5 of the filler items used the polar prompt and 5 used the scalar prompt. The results were not analyzed. The comic strip contexts are identical to those used in Experiment 1. See Appendix A for comics.

Item 7

Title	At a party
Antecedent	<i>Host:</i> Do you want another drink?
Reply	<i>Guest:</i> I really can't do that.
Polar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the guest meant: I can't have another drink of any kind.
Scalar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the guest meant: I can't have another cocktail, but maybe a beer would be OK.

Item 8

Title	At a party
Antecedent	<i>Host:</i> Do you want a big piece of cake?
Reply	<i>Guest:</i> I shouldn't do that.
Polar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the guest meant: I shouldn't have any cake at all.
Scalar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the guest meant: I shouldn't have a big piece of cake, but maybe a smaller piece would be OK.

Item 9

Title	At a car dealership
Antecedent	<i>Salesperson:</i> I see you're looking at the fast car.
Reply	<i>Wife:</i> We can't do that.
Polar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the wife meant: We can't buy a car at all today.
Scalar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the wife meant: We can't buy a fast car today, but we might be willing to buy a slower car.

Item 10

Title	In a jewelry store
Antecedent	<i>Salesperson:</i> Were you looking to buy a ring today?
Reply	<i>Customer:</i> I can't do that.
Polar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the customer meant: I can't buy a ring at all today.

Scalar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the customer meant: I can't buy an expensive ring today, but a cheaper one might be OK.
Item 11	
Title	In a liquor store
Antecedent	<i>Man:</i> Should we buy the expensive wine?
Reply	<i>Woman:</i> We shouldn't do that.
Polar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the woman meant: We shouldn't buy any wine today.
Scalar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the woman meant: We shouldn't buy the expensive wine today, but maybe a cheaper wine would be OK.
Item 12	
Title	At a buffet
Statement	<i>Father:</i> Don't do that!
Polar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the father meant: Don't take any of that.
Scalar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the father meant: Don't take so much of that.
Item 13	
Title	At a bake sale
Antecedent	<i>Woman:</i> I'm going to buy a whole pie!
Reply	<i>Man:</i> Don't do that!
Polar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the man meant: Don't buy any pie at all.
Scalar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the man meant: Don't buy a whole pie, but maybe a slice of pie would be OK.
Item 14	
Title	Opening a lock
Statement	<i>Standing person:</i> You can't do that.
Polar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the standing person meant: You can't open that lock at all.

Scalar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the standing person meant: You can't open that lock with that key.
Item 15	
Title	Turning a screw
Statement	<i>Standing person:</i> You can't do that.
Polar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the standing person meant: You can't turn that screw at all.
Scalar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the standing person meant: You can't turn that screw with that screwdriver.
Item 16	
Title	Writing
Antecedent	<i>Person on right:</i> Can you hand me the short pencil?
Reply	<i>Person on left:</i> I can't do that.
Polar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the person on the left meant: I can't hand you any pencil at all.
Scalar prompt	On a scale of 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the person on the left meant: I can't hand you the short pencil, but maybe I can hand you a longer pencil.

APPENDIX E

EXPERIMENT 5: EMPIRICAL PRIORS, THREE MESSAGES

E.1 Task and layout

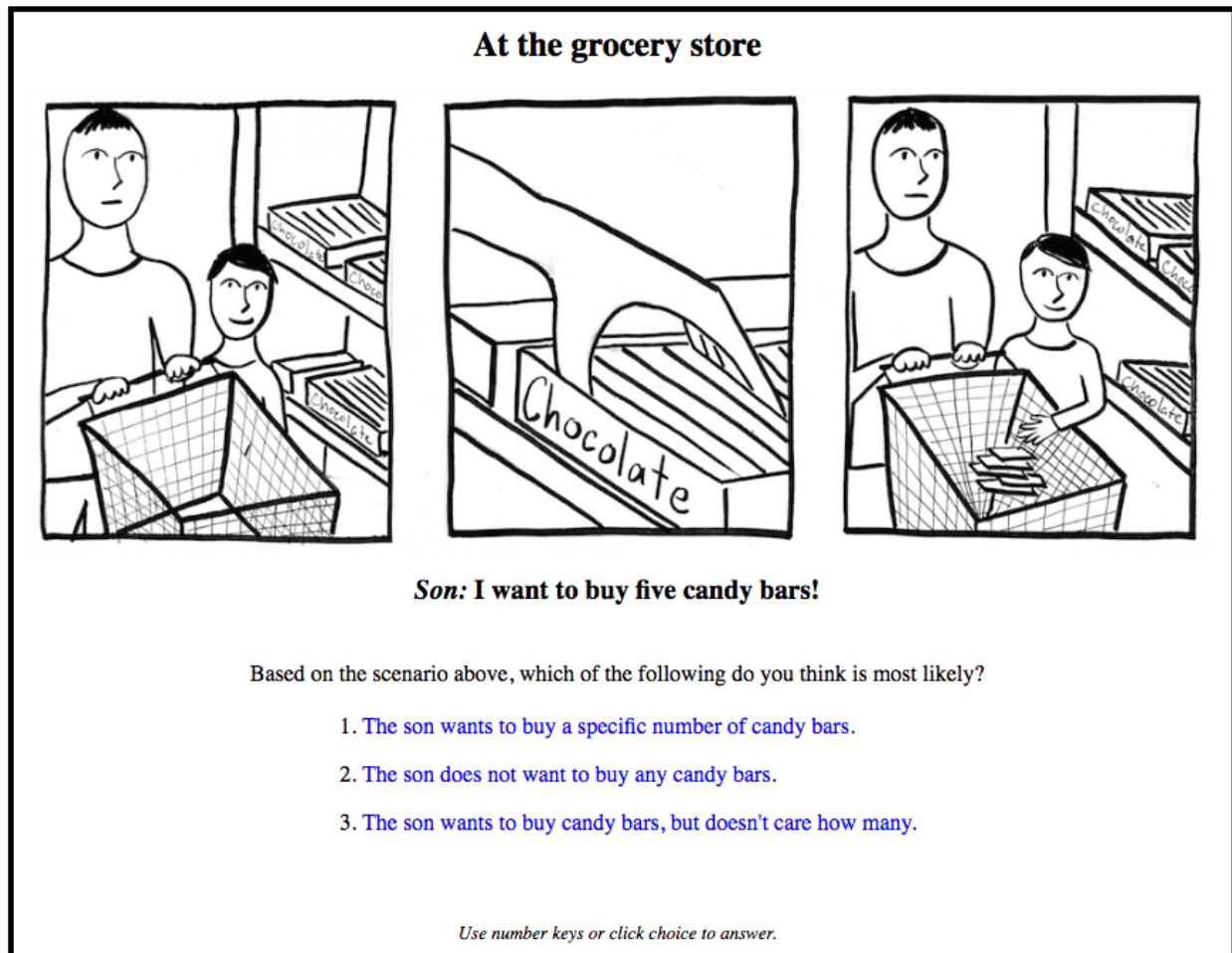


Figure E.1: Sample screen from Experiment 5

E.2 Participant statistics and data loss

Participants:	227
Gender:	103 female, 124 male
Mean age:	34.8 years
Non-native speakers excluded:	2 (0.9%)
Attention criterion 1:	Participant’s mean reaction time $\leq$ 2000 ms
Inattentive participants excluded:	14 (6.2%)
Attention criterion 2:	2 or 3 incorrect responses on 3 controlled fillers
Additional inattentive participants excluded:	10 (4.4%)
Average completion time (included participants):	7 min, 12 sec
Payment:	USD 1.00

E.3 Critical items

The comics for each item were the same as in Experiment 1. See Appendix A for comics.

Item 1

Title	At the grocery store
Unmodified antecedent	<i>Son</i> I want to buy candy bars!
Modified antecedent	<i>Son</i> I want to buy five candy bars!
Prompt	Based on the scenario above, which of the following do you think is most likely?
Zero choice	The son does not want to buy any candy bars.
Indifferent choice	The son wants to buy candy bars, but doesn’t care how many.
Specific choice	The son wants to buy a specific number of candy bars.

Item 2

Title	In an office
Unmodified antecedent	<i>Standing worker</i> I’m going to take candies!
Modified antecedent	<i>Standing worker</i> I’m going to take four candies!
Prompt	Based on the scenario above, which of the following do you think is most likely?
Zero choice	The standing worker does not want any candy.
Indifferent choice	The standing worker wants candy, but he doesn’t care how many pieces.
Specific choice	The standing worker wants a specific number of pieces of candy.

Item 3

Title	In a cooking class
Unmodified antecedent	<i>Student</i> Now I'm going to add chili peppers.
Modified antecedent	<i>Student</i> Now I'm going to add four chili peppers.
Prompt	Based on the scenario above, which of the following do you think is most likely?
Zero choice	The student is not planning to add any peppers.
Indifferent choice	The student is planning to add peppers, but she doesn't care how many.
Specific choice	The student is planning to add a specific number of peppers.

Item 4

Title	In a jewelry store
Unmodified antecedent	<i>Wife</i> I want to get charms for my bracelet.
Modified antecedent	<i>Wife</i> I want to get four charms for my bracelet.
Prompt	Based on the scenario above, which of the following do you think is most likely?
Zero choice	The wife does not want to buy any charms.
Indifferent choice	The wife wants to buy some charms, but she doesn't care how many.
Specific choice	The wife wants to buy a specific number of charms.

Item 5

Title	In a clothing store
Unmodified antecedent	<i>Salesperson</i> I think you should buy shirts today.
Modified antecedent	<i>Salesperson</i> I think you should buy three shirts today.
Prompt	Based on the scenario above, which of the following do you think is most likely?
Zero choice	The clerk doesn't think the man should buy any shirts.
Indifferent choice	The clerk thinks the man should buy shirts, but doesn't care how many.
Specific choice	The clerk thinks the man should buy a specific number of shirts.

Item 6

Title	In a children's store
Unmodified antecedent	<i>Daughter</i> I want to buy princess dresses!
Modified antecedent	<i>Daughter</i> I want to buy three princess dresses!
Prompt	Based on the scenario above, which of the following do you think is most likely?
Zero choice	The daughter does not want to buy any dresses.

Indifferent choice

The daughter wants to buy dresses, but she doesn't care how many.

Specific choice

The daughter wants to buy a specific number of dresses.

E.4 Fillers – Not analyzed

The following filler items were not analyzed.

Item 7

Title

At a party

Unavailable comic



Prompt

Based on the scenario above, which of the following do you think is most likely?

Choice 1

The man will accept the martini.

Choice 2

The man won't take a drink.

Choice 3

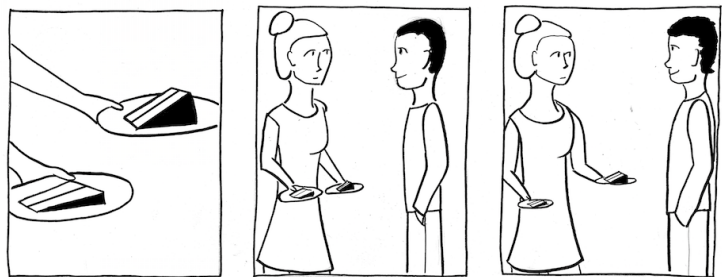
The man will ask for a beer.

Item 8

Title

At a party

Unavailable comic



Antecedent

Host: Would you like another drink?

Prompt

Based on the scenario above, which of the following do you think is most likely?

Choice 1

The man will take a small piece of cake.

Choice 2

The man will ask for a big piece of cake.

Choice 3

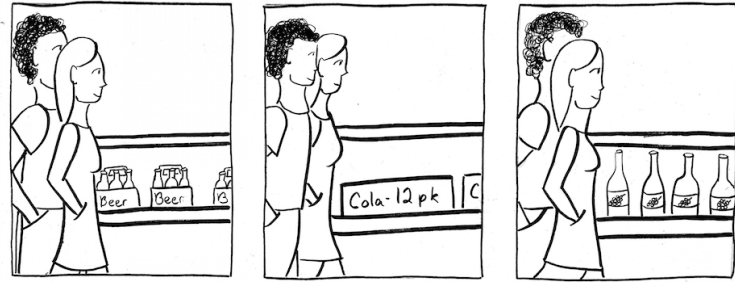
The man won't take any cake.

Item 9

Title

In a liquor store

Unavailable comic



Prompt

Based on the scenario above, which of the following do you think is most likely?

Choice 1

The couple will buy beer.

Choice 2

The couple will buy cola.

Choice 3

The couple will buy wine.

E.5 Fillers – Attention screening

The following filler items were used to screen for participant attention. Each was annotated with a correct response and the data from participants answering two or three items incorrectly was excluded from analysis.

Item 10

Title

At a bake sale

Unavailable comic



Prompt

Based on the scenario above, which of the following do you think is most likely?

Correct response

The woman picked up a pie.

Incorrect response 1

The woman picked up a bag of cookies.

Incorrect response 2

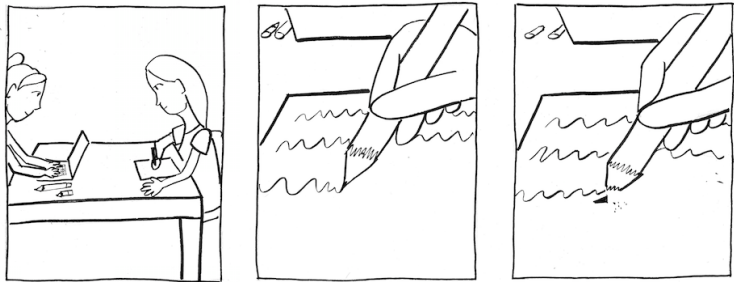
The woman did not pick anything up.

Item 11

Title

Writing

Unavailable comic



Antecedent

Person on right: Oops, I broke my pencil.

Prompt

Based on the scenario above, which of the following do you think is most likely?

Correct response

The woman broke the tip of a pencil.

Incorrect response 1

The woman broke a pen.

Incorrect response 2

The woman broke a table.

Item 12

Title

Opening a lock

Unavailable comic



Prompt

Based on the scenario above, which of the following do you think is most likely?

Correct response

The woman could not open the lock.

Incorrect response 1

The woman opened the lock on the first try.

Incorrect response 2

The woman successfully opened the lock after several tries.

APPENDIX F

EXPERIMENT 6: EMPIRICAL PRIORS, TWO MESSAGES

F.1 Task and layout

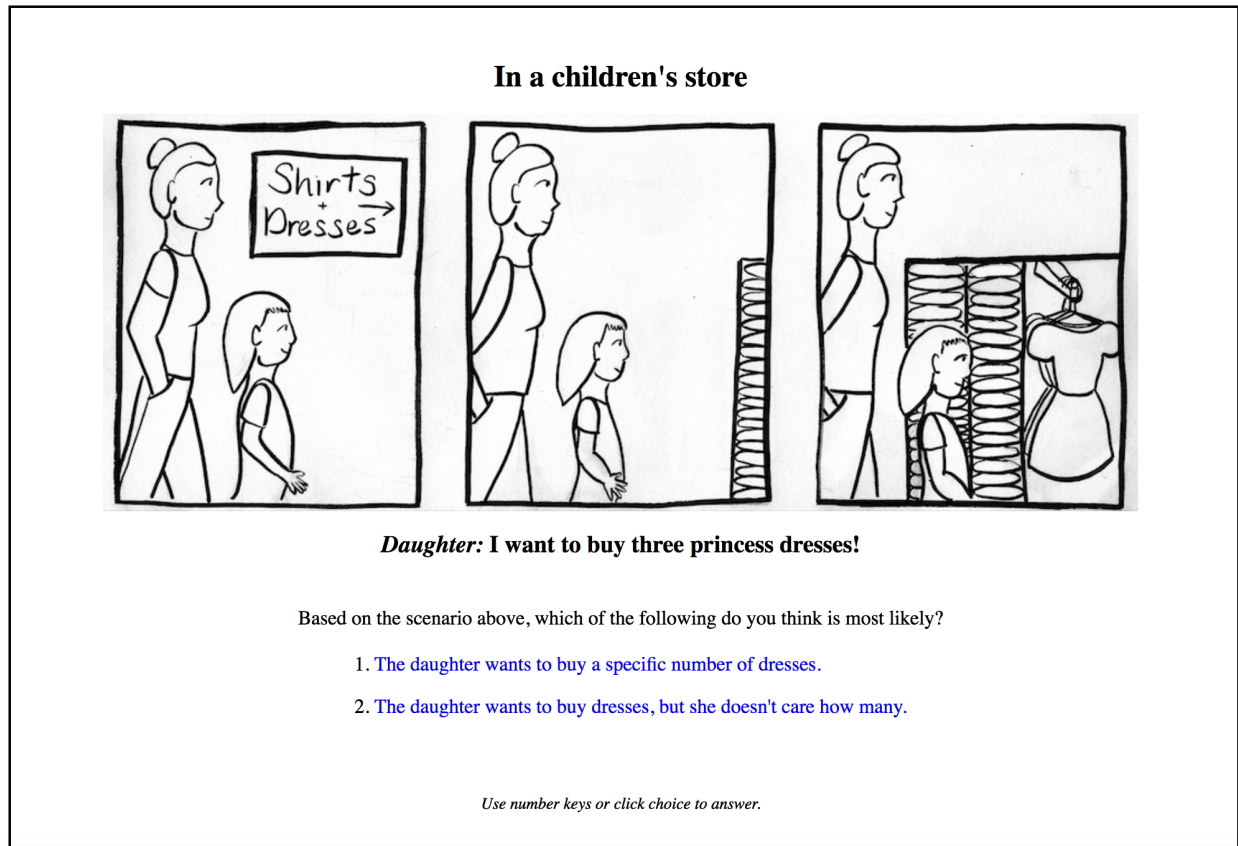


Figure F.1: Sample screen from Experiment 6

F.2 Participant statistics and data loss

Participants:	224
Gender:	111 female, 113 male
Mean age:	35.6 years
Non-native speakers excluded:	6 (2.7%)
Attention criterion 1:	Participant's mean reaction time $\leq$ 2000 ms
Additional inattentive participants excluded:	4 (1.8%)
Attention criterion 2:	2 or 3 incorrect responses on 3 controlled fillers
Additional inattentive participants excluded:	5 (2.2%)
Average completion time (included participants):	6 min, 18 sec
Payment:	USD 1.00

F.3 Critical items

The critical items in Experiment 6 were identical to those from Experiment 5 except that the “zero” choice was not available for selection. See Appendix E for materials.

F.4 Fillers

The fillers and their use to screen participants were identical to Experiment 5, except that the following choices were removed from the corresponding items to match the format of the critical trials. The choices removed from items 10-12 were incorrect responses.

- Item 7** The man won't take a drink.
- Item 8** The man won't take any cake.
- Item 9** The couple will buy beer.
- Item 10** The woman did not pick anything up.
- Item 11** The woman broke a pen.
- Item 12** The woman successfully opened the lock after several tries.

APPENDIX G

EXPERIMENT 7: FORCED-CHOICE VPE INTERPRETATION (EMPIRICAL POSTERIORS)

G.1 Task and layout

At the grocery store



Son: I want to buy candy bars!

Father: We can't.

Based on the scenario above, do you think it is more likely that **the father** meant:

1. We can't buy five candy bars, but maybe we could buy fewer.
2. We can't buy any candy bars.

Use number keys or click choice to answer.

Figure G.1: Sample screen from Experiment 7

G.2 Participant statistics and data loss

Participants:	198
Gender:	85 female, 113 male
Mean age:	35.4 years
Non-native speakers excluded:	4 (2.0%)
Attention criterion 1:	Participant’s mean reaction time $\leq$ 2000 ms
Inattentive participants excluded:	7 (3.5%)
Attention criterion 2:	2 or 3 incorrect responses on 3 controlled fillers
Additional inattentive participants excluded:	9 (4.5%)
Average completion time (included participants):	7 min, 21 sec
Payment:	USD 1.00

G.3 Critical items

The comics used were the same as in Experiment 1. See Appendix A for comics.

Item 1

Title	At the grocery store
Unmodified antecedent	<i>Son</i> : I want to buy candy bars!
Modified antecedent	<i>Son</i> : I want to buy five candy bars!
Reply	<i>Father</i> : We can’t.
Prompt	Based on the scenario above, do you think it is more likely that the father meant:
Polar choice	We can’t buy any candy bars.
Scalar choice	We can’t buy five candy bars, but maybe we could buy fewer.

Item 2

Title	In an office
Unmodified antecedent	<i>Standing worker</i> : I’m going to take candies!
Modified antecedent	<i>Standing worker</i> : I’m going to take four candies!
Reply	<i>Sitting worker</i> : You can’t.
Prompt	Based on the scenario above, do you think it is more likely that the sitting worker meant:
Polar choice	You can’t take any candies.
Scalar choice	You can’t take four candies, but maybe you can take fewer.

Item 3

Title	In a cooking class
Unmodified antecedent	<i>Student:</i> Now I'm going to add chili peppers.
Modified antecedent	<i>Student:</i> Now I'm going to add four chili peppers.
Reply	<i>Instructor:</i> You shouldn't.
Prompt	Based on the scenario above, do you think it is more likely that the instructor meant:
Polar choice	You shouldn't add any peppers.
Scalar choice	You shouldn't add four chili peppers, but maybe fewer would be OK.

Item 4

Title	In a jewelry store
Unmodified antecedent	<i>Wife:</i> I want to get charms for my bracelet.
Modified antecedent	<i>Wife:</i> I want to get four charms for my bracelet.
Reply	<i>Husband:</i> You shouldn't.
Prompt	Based on the scenario above, do you think it is more likely that the husband meant:
Polar choice	You shouldn't buy any charms.
Scalar choice	You shouldn't buy four charms, but maybe fewer would be OK.

Item 5

Title	In a clothing store
Unmodified antecedent	<i>Salesperson:</i> I think you should buy shirts today.
Modified antecedent	<i>Salesperson:</i> I think you should buy three shirts today.
Reply	<i>Customer:</i> I shouldn't.
Prompt	Based on the scenario above, do you think it is more likely that the customer meant:
Polar choice	I shouldn't buy any shirts.
Scalar choice	I shouldn't buy three shirts, but maybe fewer would be OK.

Item 6

Title	In a children's store
Unmodified antecedent	<i>Daughter:</i> I want to buy princess dresses!
Modified antecedent	<i>Daughter:</i> I want to buy three princess dresses!
Reply	<i>Mother:</i> We can't.
Prompt	Based on the scenario above, do you think it is more likely that the mother meant:
Polar choice	We can't buy any dresses.
Scalar choice	We can't buy three dresses, but maybe we can buy fewer.

G.4 Fillers – Not analyzed

The following fillers were not analyzed.

Item 7

Title

At a party

Comic



Antecedent

Host: Do you want another drink?

Reply

Guest: I really can't.

Prompt

Based on the scenario above, do you think it is more likely that **the guest** meant:

Choice 1

I can't have another drink of any kind.

Choice 2

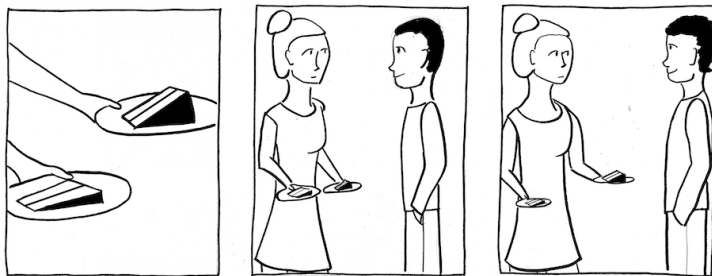
I can't have another cocktail, but maybe a beer would be OK.

Item 8

Title

At a party

Comic



Antecedent

Host: Do you want a big piece of cake?

Reply

Guest: I shouldn't.

Prompt

Based on the scenario above, do you think it is more likely that **the guest** meant:

Choice 1

I shouldn't have any cake at all.

Choice 2

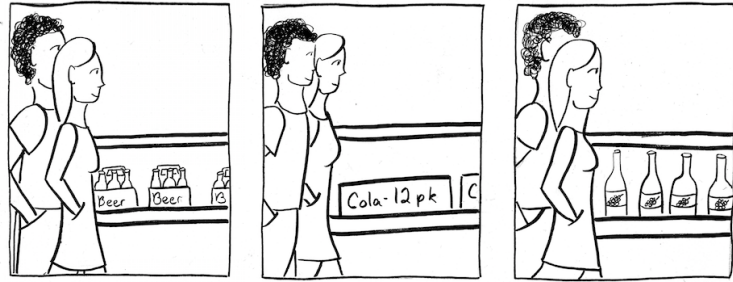
I shouldn't have a big piece of cake, but maybe a smaller piece would be OK.

Item 9

Title

In a liquor store

Comic



Antecedent

Man: Should we buy the expensive wine?

Reply

Woman: We shouldn't.

Prompt

Based on the scenario above, do you think it is more likely that **the woman** meant:

Choice 1

We shouldn't buy any wine.

Choice 2

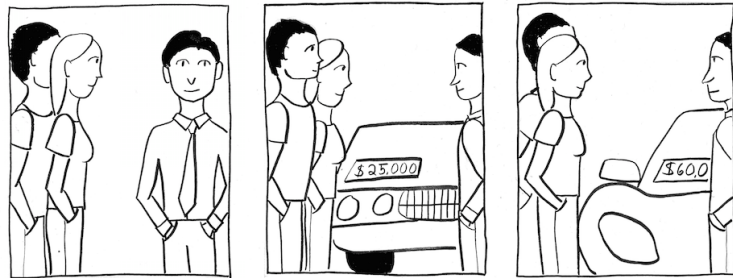
We shouldn't buy the expensive wine, but maybe a cheaper wine would be OK.

Item 10

Title

At a car dealership

Comic



Antecedent

Salesperson: I see you're looking at the fast car.

Reply

Wife: We can't.

Prompt

Based on the scenario above, do you think it is more likely that **the wife** meant:

Choice 1

We can't buy a car at all.

Choice 2

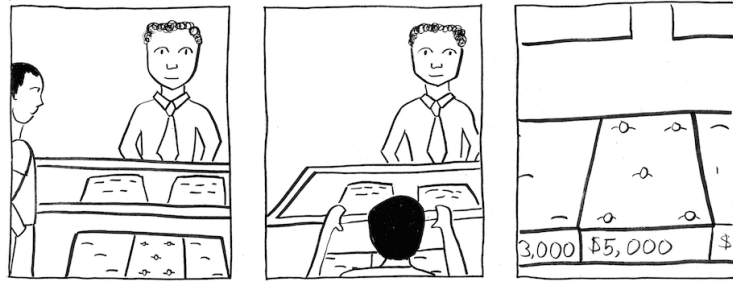
We can't buy a fast car, but maybe a slower car would be OK.

Item 11

Title

In a jewelry store

Comic



Antecedent

Salesperson: Were you looking to buy a ring today?

Reply

Customer: I can't.

Prompt

Based on the scenario above, do you think it is more likely that **the customer** meant:

Choice 1

I can't buy a ring at all.

Choice 2

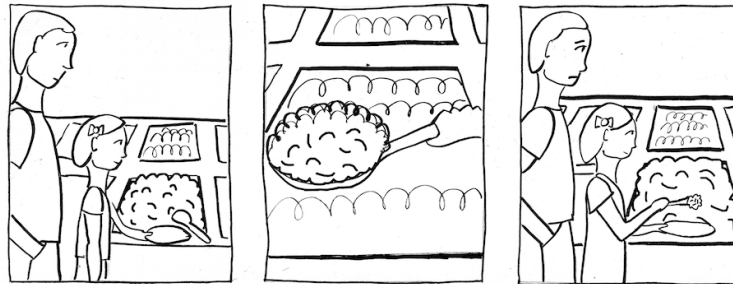
I can't buy an expensive ring, but maybe a cheaper ring would be OK.

Item 12

Title

At a buffet

Comic



Statement

Father: Don't!

Prompt

Based on the scenario above, do you think it is more likely that **the father** meant:

Choice 1

Don't take any of that.

Choice 2

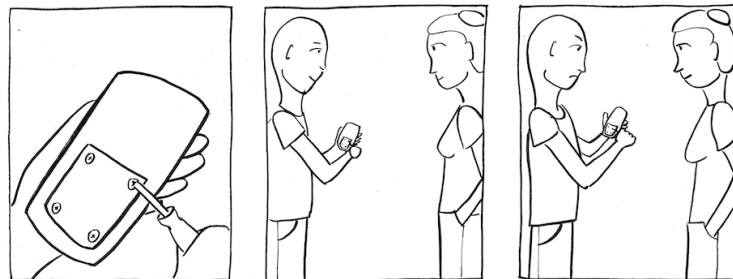
Don't take so much of that.

Item 13

Title

Turning a screw

Comic



Statement

Standing person: You can't.

Prompt

Based on the scenario above, do you think it is more likely that **the standing person** meant:

Choice 1

You can't turn that screw at all.

Choice 2

You can't turn that screw with that screwdriver.

G.5 Fillers – Attention screening

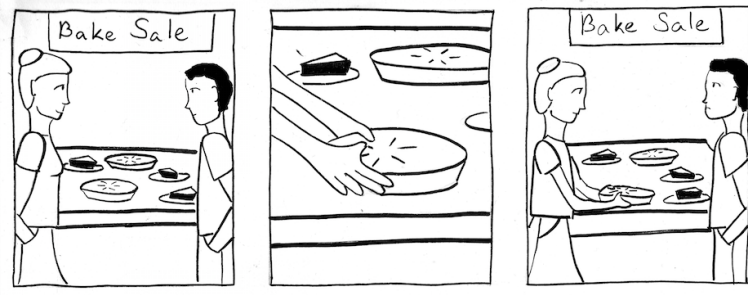
The following filler items were used to screen for participant attention. Each was annotated with a correct response and the data from participants answering two or three items incorrectly was excluded from analysis.

Item 11

Title

At a bake sale

Comic



Antecedent

Woman: I'm going to buy a whole pie.

Reply

Man: Don't!

Prompt

Based on the scenario above, do you think it is more likely that **the man** meant:

Correct choice

Don't buy a whole pie.

Incorrect choice

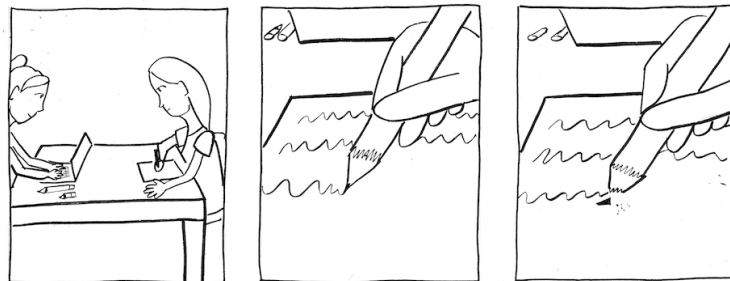
Don't bake a whole pie.

Item 12

Title

Writing

Comic



Antecedent

Person on right: Can you hand me another pencil?

Reply

Person on left: I can't!

Prompt
Correct choice
Incorrect choice

Based on the scenario above, do you think it is more likely that **the person on the left** meant:
I can't hand you a pencil.
I can't hand you a notebook.

Item 13
Title

Opening a lock

Comic



Statement

Standing person: You can't!

Prompt

Based on the scenario above, do you think it is more likely that **the person standing up** meant:

Correct choice
Incorrect choice

You can't open the lock with that key.
You can't open the door with that key.

APPENDIX H

MODELING CALCULATIONS AND MODEL COMPARISON

H.1 Sample model calculations

This section shows sample calculations for each of the models considered in Chapter 4. Each equation will show the calculation of the predicted posterior probability of concluding the polar or scalar interpretation in the unmodified/salient condition. Recall that in this condition, the discourse strategy assigns a truth value/compatibility rating of 0 for the polar interpretation and 1 for the scalar interpretation, while the salience strategy assigns a rating of 1 for both. The prior probability of the polar interpretation as estimated by Experiment 5 with the zero responses excluded is .240, while the prior probability of the scalar interpretation is .760. For the purpose of illustration, the models with free parameters will use $\epsilon = .05$ and/or $\beta = .55$. In the equations below, *VPE* will stand for the VPE reply as *u*, *us* will represent the context of utterance as *c* (unmodified antecedent and salient context), and *p* and *s* will be used to represent the polar and scalar interpretations, respectively as *m*. For reference, the empirical posterior probabilities for the polar and scalar interpretations as measured in Experiment 7 are .567 and .433, respectively.

Identity model without error, polar

$$\begin{aligned}
 P(m|u, c) &= \frac{\llbracket u \rrbracket_{\text{identity}}^{m,c} \cdot P(m|c)}{\sum_{m' \in M} \llbracket u \rrbracket_{\text{identity}}^{m',c} \cdot P(m'|c)} \\
 P(p|VPE, us) &= \frac{\llbracket VPE \rrbracket_{\text{identity}}^{p, us} \cdot P(p|us)}{\llbracket VPE \rrbracket_{\text{identity}}^{p, us} \cdot P(p|us) + \llbracket VPE \rrbracket_{\text{identity}}^{s, us} \cdot P(s|us)} \\
 &= \frac{1 \cdot .240}{1 \cdot .240 + 0 \cdot .760} \\
 &= 1
 \end{aligned}$$

Identity model without error, scalar

$$\begin{aligned}
 P(m|u, c) &= \frac{\llbracket u \rrbracket_{\text{identity}}^{m,c} \cdot P(m|c)}{\sum_{m' \in M} \llbracket u \rrbracket_{\text{identity}}^{m',c} \cdot P(m'|c)} \\
 P(s|\text{VPE}, \text{us}) &= \frac{\llbracket \text{VPE} \rrbracket_{\text{identity}}^{\text{s}, \text{us}} \cdot P(s|\text{us})}{\llbracket \text{VPE} \rrbracket_{\text{identity}}^{\text{s}, \text{us}} \cdot P(s|\text{us}) + \llbracket \text{VPE} \rrbracket_{\text{identity}}^{\text{p}, \text{us}} \cdot P(p|\text{us})} \\
 &= \frac{0 \cdot .760}{1 \cdot .760 + 1 \cdot .240} \\
 &= 0
 \end{aligned}$$

Identity model with error, polar

$$\begin{aligned}
 P(m|u, c) &= (1 - \epsilon) \frac{\llbracket u \rrbracket_{\text{identity}}^{m,c} \cdot P(m|c)}{\sum_{m' \in M} \llbracket u \rrbracket_{\text{identity}}^{m',c} \cdot P(m'|c)} + \epsilon \frac{1}{|M|} \\
 P(p|\text{VPE}, \text{us}) &= (1 - \epsilon) \frac{\llbracket \text{VPE} \rrbracket_{\text{identity}}^{\text{p}, \text{us}} \cdot P(p|\text{us})}{\llbracket \text{VPE} \rrbracket_{\text{identity}}^{\text{p}, \text{us}} \cdot P(p|\text{us}) + \llbracket \text{VPE} \rrbracket_{\text{identity}}^{\text{s}, \text{us}} \cdot P(s|\text{us})} + \epsilon \frac{1}{|M|} \\
 &= (1 - .05) \frac{1 \cdot .240}{1 \cdot .240 + 0 \cdot .760} + (.05) \frac{1}{2} \\
 &= .975
 \end{aligned}$$

Identity model with error, scalar

$$\begin{aligned}
 P(m|u, c) &= (1 - \epsilon) \frac{\llbracket u \rrbracket_{\text{identity}}^{m,c} \cdot P(m|c)}{\sum_{m' \in M} \llbracket u \rrbracket_{\text{identity}}^{m',c} \cdot P(m'|c)} + \epsilon \frac{1}{|M|} \\
 P(s|\text{VPE}, \text{us}) &= (1 - \epsilon) \frac{\llbracket \text{VPE} \rrbracket_{\text{identity}}^{\text{s}, \text{us}} \cdot P(s|\text{us})}{\llbracket \text{VPE} \rrbracket_{\text{identity}}^{\text{s}, \text{us}} \cdot P(s|\text{us}) + \llbracket \text{VPE} \rrbracket_{\text{identity}}^{\text{p}, \text{us}} \cdot P(p|\text{us})} + \epsilon \frac{1}{|M|} \\
 &= (1 - .05) \frac{0 \cdot .760}{0 \cdot .760 + 1 \cdot .240} + (.05) \frac{1}{2} \\
 &= .025
 \end{aligned}$$

Saliency model without error, polar

Note that the saliency and hybrid model calculations use the simplified saliency formula. Recall that since the saliency strategy assigns a rating of 1 to all interpretations, the denominator of the unsimplified model is always 1 and can be ignored.

$$\begin{aligned}P(m|u, c) &= P(m|c) \\P(p|VPE, us) &= P(p|us) \\&= .240\end{aligned}$$

Saliency model without error, scalar

$$\begin{aligned}P(m|u, c) &= P(m|c) \\P(s|VPE, us) &= P(s|us) \\&= .760\end{aligned}$$

Saliency model with error, polar

$$\begin{aligned}P(m|u, c) &= (1 - \epsilon) \cdot P(m|c) + \epsilon \frac{1}{|M|} \\P(p|VPE, us) &= (1 - \epsilon) \cdot P(p|us) + \epsilon \frac{1}{|M|} \\&= (1 - .05) \cdot .240 + (.05) \frac{1}{2} \\&= .253\end{aligned}$$

Saliience model with error, scalar

$$\begin{aligned}
 P(m|u, c) &= (1 - \epsilon) \cdot P(m|c) + \epsilon \frac{1}{|M|} \\
 P(s|VPE, us) &= (1 - \epsilon) \cdot P(s|us) + \epsilon \frac{1}{|M|} \\
 &= (1 - .05) \cdot .760 + (.05) \frac{1}{2} \\
 &= .747
 \end{aligned}$$

Hybrid model without error, polar

$$\begin{aligned}
 P(m|u, c) &= \beta \cdot \frac{\llbracket u \rrbracket_{\text{identity}}^{m,c} \cdot P(m|c)}{\sum_{m' \in M} \llbracket u \rrbracket_{\text{identity}}^{m',c} \cdot P(m'|c)} + (1 - \beta) \cdot P(m|c) \\
 P(p|VPE, us) &= \beta \cdot \frac{\llbracket VPE \rrbracket_{\text{identity}}^{p,us} \cdot P(p|us)}{\llbracket VPE \rrbracket_{\text{identity}}^{p,us} \cdot P(p|us) + \llbracket VPE \rrbracket_{\text{identity}}^{s,us} \cdot P(s|us)} + (1 - \beta) \cdot P(p|us) \\
 &= .55 \cdot \frac{1 \cdot .240}{1 \cdot .240 + 0 \cdot .760} + (1 - .55) \cdot .240 \\
 &= .658
 \end{aligned}$$

Hybrid model without error, scalar

$$\begin{aligned}
 P(m|u, c) &= \beta \cdot \frac{\llbracket u \rrbracket_{\text{identity}}^{m,c} \cdot P(m|c)}{\sum_{m' \in M} \llbracket u \rrbracket_{\text{identity}}^{m',c} \cdot P(m'|c)} + (1 - \beta) \cdot P(m|c) \\
 P(s|VPE, us) &= \beta \cdot \frac{\llbracket VPE \rrbracket_{\text{identity}}^{s,us} \cdot P(s|us)}{\llbracket VPE \rrbracket_{\text{identity}}^{s,us} \cdot P(s|us) + \llbracket VPE \rrbracket_{\text{identity}}^{p,us} \cdot P(p|us)} + (1 - \beta) \cdot P(s|us) \\
 &= .55 \cdot \frac{0 \cdot .760}{0 \cdot .760 + 1 \cdot .240} + (1 - .55) \cdot .760 \\
 &= .342
 \end{aligned}$$

Hybrid model with error, polar

$$\begin{aligned}
P(m|u, c) &= (1 - \epsilon) \left(\beta \cdot \frac{\llbracket u \rrbracket_{\text{identity}}^{m,c} \cdot P(m|c)}{\sum_{m' \in M} \llbracket u \rrbracket_{\text{identity}}^{m',c} \cdot P(m'|c)} + (1 - \beta) \cdot P(m|c) \right) + \epsilon \cdot \frac{1}{|M|} \\
P(p|us) &= (1 - \epsilon) \left(\beta \cdot \frac{\llbracket \text{VPE} \rrbracket_{\text{identity}}^{\text{p},us} \cdot P(p|us)}{\llbracket \text{VPE} \rrbracket_{\text{identity}}^{\text{p},us} \cdot P(p|us) + \llbracket \text{VPE} \rrbracket_{\text{identity}}^{\text{s},us} \cdot P(s|us)} + (1 - \beta) \cdot P(p|us) \right) + \epsilon \cdot \frac{1}{|M|} \\
&= (1 - .05) \left(.55 \cdot \frac{1 \cdot .240}{1 \cdot .240 + 0 \cdot .760} + (1 - .55) \cdot .240 \right) + .05 \cdot \frac{1}{2} \\
&= .650
\end{aligned}$$

Hybrid model with error, scalar

$$\begin{aligned}
P(m|u, c) &= (1 - \epsilon) \left(\beta \cdot \frac{\llbracket u \rrbracket_{\text{identity}}^{m,c} \cdot P(m|c)}{\sum_{m' \in M} \llbracket u \rrbracket_{\text{identity}}^{m',c} \cdot P(m'|c)} + (1 - \beta) \cdot P(m|c) \right) + \epsilon \cdot \frac{1}{|M|} \\
P(s|us) &= (1 - \epsilon) \left(\beta \cdot \frac{\llbracket \text{VPE} \rrbracket_{\text{identity}}^{\text{s},us} \cdot P(s|us)}{\llbracket \text{VPE} \rrbracket_{\text{identity}}^{\text{s},us} \cdot P(s|us) + \llbracket \text{VPE} \rrbracket_{\text{identity}}^{\text{p},us} \cdot P(p|us)} + (1 - \beta) \cdot P(s|us) \right) + \epsilon \cdot \frac{1}{|M|} \\
&= (1 - .05) \left(.55 \cdot \frac{0 \cdot .760}{0 \cdot .760 + 1 \cdot .240} + (1 - .55) \cdot .760 \right) + .05 \cdot \frac{1}{2} \\
&= .350
\end{aligned}$$

H.2 Bayes factors comparing identity and salience models

The tables below give the Bayes factor by context/antecedent combination for the pairs of models that were not given in Chapter 4, namely each pairwise comparison of identity and salience models with and without error. Note that each table shows progressively fewer comparisons, as the value comparing one model in the numerator with another value in the denominator is the reciprocal of the opposite comparison; thus, all of the possible comparisons are represented in the tables in Chapter 4 or the tables below.

Saliency model with error as Model 1

<i>Antecedent</i>	<i>Context</i>	<i>Model 2</i>		
		Identity w/o ϵ	Identity w/ ϵ	Saliency w/o ϵ
Exophoric	Unavailable	0.717	1.00	0.717
Exophoric	Available	1.72	1.00	1.72
Exophoric	Salient	0.489	1.00	0.489
Unmodified	Unavailable	∞	0.361	0.305
Unmodified	Available	∞	1.90e03	0.437
Unmodified	Salient	∞	3.85e11	143
Modified	Unavailable	2.28	1.00	2.28
Modified	Available	1.31	1.00	1.31
Modified	Salient	0.854	1.00	0.854

Table H.1: Bayes factors by condition with saliency model with error as Model 1.

Saliency model without error as Model 1

<i>Antecedent</i>	<i>Context</i>	<i>Model 2</i>	
		Identity w/o ϵ	Identity w/ ϵ
Exophoric	Unavailable	1.00	1.39
Exophoric	Available	1.00	0.583
Exophoric	Salient	1.00	2.04
Unmodified	Unavailable	∞	1.18
Unmodified	Available	∞	4.36e03
Unmodified	Salient	∞	2.70e09
Modified	Unavailable	1.00	0.439
Modified	Available	1.00	0.766
Modified	Salient	1.00	1.17

Table H.2: Bayes factors by condition with discourse model without error as Model 1.

Identity model with error as Model 1

	<i>Antecedent</i>	<i>Context</i>		
		Unavailable	Available	Salient
	Exophoric	0.717	1.72	0.489
	Unmodified	∞	∞	∞
	Modified	2.28	1.31	0.854

Table H.3: Bayes factors by condition with identity model with error as Model 1 and identity model without error as Model 2.

APPENDIX I

EXPERIMENT 8: NORMING VERB INFERABILITY

I.1 Note on “new” verbs

One possible issue with the new-condition verbs used in the deaccenting chapters is that some “new” pairs, e.g., *dressed-harassed*, are merely semantically unrelated to one another, while other pairs, e.g., *rejected-endorsed*, are closer to antonyms. The concern is that the “new” second-clause verbs might not behave homogeneously because of this contrast; verbs that are antonyms of the first-clause verb might be more likely to be realized with focus/prosody intended to highlight the contrast between the two verbs, while verbs that are merely unrelated might have a categorically different focus/prosodic realization (i.e., the verbs are merely “new”).

While this lack of homogeneity in the “new” verbs was unintentional, it is not likely to have affected the findings for Experiments 9 and 10. The new condition acts primarily as a baseline against which to compare the realization of the inference conditions. New verbs canonically should be accented, and comparing the inference conditions to them is a way to determine whether the realization of inference verbs is significantly different from that of canonically accented verbs.

The accidental inclusion of some “new” verbs that might receive contrastive focus is thus actually conservative. This is because the realization of contrastive focus should be, if anything, stronger (i.e., higher phonetic correlates) than the realization of merely “new” constituents. Since the primary question of the experiment is whether the phonetic correlates of inference verbs are lower than those of new verbs, then, the inclusion of some items that may unintentionally increase the mean values for new verbs does not detract from the conclusion that the means for inference verbs were not lower than the means for new verbs.

I.2 Task and layout

Use the scale to answer the question, where 7 represents *very likely* and 1 represents *very unlikely*.

Suppose you know that

Raymond despised Gregory.

How likely do you think it is that

Raymond hated Gregory?

(Unlikely) (Likely)

Use number keys or click boxes to answer.

Figure I.1: Sample screen from Experiment 8

I.3 Participant statistics and data loss

Participants:	60
Gender:	36 male, 24 female
Mean age:	33.2 years
Non-native speakers excluded:	0 (0%)
Attention criterion:	Participant's mean reaction time $\leq$ 2000ms
Inattentive participants excluded:	5 (8.3%)
Average completion time (included participants):	10 minutes, 16 seconds
Payment:	USD 1.00

I.4 Aggregate results

Relation	Condition	Mean
Entailment	Inference	6.49
Entailment	New	2.47
Bridging	Inference	5.84
Bridging	New	2.43

Table I.1: Mean verb inferability scores by relation type and condition, all items

Relation	Condition	Mean
Entailment	Inference	6.49
Entailment	New	2.12
Bridging	Inference	5.78
Bridging	New	2.15

Table I.2: Mean verb inferability scores, perception items (24 entailment and 24 bridging items used in Chapter 7)

Relation	Condition	Mean
Entailment	Inference	6.56
Entailment	New	1.86
Bridging	Inference	5.49
Bridging	New	2.19

Table I.3: Mean verb inferability scores, production items (6 entailment and 6 bridging items used in Chapter 6)

I.5 By-item results

	Antecedent	Inferencing target	Inferencing mean	New target	New mean
	consoled	comforted	6.43	applauded	2.90
†	neglected	ignored	6.33	berated	4.13
	scammed	cheated	6.70	affirmed	2.05
*	entertained	amused	6.38	delayed	3.07
†	bested	beat	6.17	replaced	4.48
†	shocked	surprised	6.43	educated	3.33
	billed	charged	6.57	embarrassed	2.76
	scolded	admonished	6.00	feared	2.40
*	aggravated	annoyed	6.50	guided	1.48
	woke	roused	6.19	impressed	2.80
*	hugged	embraced	6.77	rebuffed	1.67
	despised	hated	6.81	helped	1.33
†	shoved	pushed	6.67	spotted	3.24
	terrified	scared	6.90	pleased	1.60
*	astounded	surprised	6.27	punished	1.95
	perveted	corrupted	6.24	sheltered	2.23
	tainted	contaminated	6.37	served	1.86
	sabotaged	undermined	6.29	fed	1.77
	captured	apprehended	6.53	strengthened	2.24
†	pestered	hassled	6.76	stopped	3.97
	released	freed	6.67	scared	2.62
	dodged	evaded	6.52	supported	1.80
	pacified	quieted	6.00	surprised	2.90
	soothed	calmed	6.81	taunted	1.57
	detested	hated	6.57	contacted	2.00
*	idolized	admired	6.81	refused	1.43
*	recommended	endorsed	6.63	rejected	1.24
†	captivated	intrigued	6.57	imagined	3.57
	floored	astonished	6.07	picked	2.71
	watched	saw	6.95	replaced	2.77

Table I.4: Entailment items: Mean verb inferability score by item

	Antecedent	Inferencing target	Inferencing mean	New target	New mean
	congratulated	complimented	6.23	upset	1.43
†	suspected	distrusted	6.14	called	3.40
	bullied	intimidated	6.53	accepted	1.38
†	alarmed	frightened	6.43	directed	3.23
	frustrated	discouraged	5.10	echoed	2.14
†	misled	tricked	6.67	battled	3.57
*	annoyed	harassed	5.67	dressed	1.76
	indulged	spoiled	5.95	threatened	1.90
	reminded	nagged	4.87	drew	2.19
†	amazed	bewildered	5.24	healed	3.80
	refuted	corrected	5.23	included	1.71
	supported	defended	6.10	duped	1.50
	ignored	excluded	5.80	instructed	1.29
	interrupted	heckled	5.29	encouraged	2.37
	placated	calmed	5.87	fetches	2.57
	detained	arrested	6.19	ignored	2.17
	fascinated	dazzled	6.10	grilled	2.14
*	charmed	seduced	5.48	offended	1.63
	subdued	overpowered	6.23	telephoned	2.10
*	misinformed	deceived	6.43	protected	2.60
	influenced	manipulated	4.97	rescued	3.00
†	outdid	outshined	6.86	questioned	3.10
*	caught	exposed	5.27	upstaged	2.62
*	tolerated	condoned	4.52	released	3.03
	adored	loved	6.40	represented	3.10
	praised	flattered	6.19	concerned	2.47
†	respected	revered	5.77	warned	3.38
*	avoided	disliked	5.62	involved	1.47
	fooled	conned	6.63	described	1.95
	approached	greeted	6.00	controlled	2.83

Table I.5: Implicational bridging items: Mean verb inferability score by item

† = items excluded after Experiment 8

* = items included in production study Experiment 9 (iambic inference targets; *admired*)

APPENDIX J

EXPERIMENT 9: PRODUCTION OF DEACCENTING

J.1 Participant statistics and data loss

Participants:	10
Gender:	5 female, 5 male
Mean age:	21.9 years
Average completion time:	34 min, 37 sec
Compensation:	USD 10.00 or course credit

Due to experimenter error, the trial for participant DA003, item 10, with an inference verb and and old object was not labeled for the forced aligner. Thus, this trial is not included in the phonetic analysis. This represents 0.1% of trials.

In addition, data for the trials shown in Table J.1 was excluded from analysis due to major speech errors identified during manual correction of the f_0 values. This represents 1.7% of the total trials. In those analyses of duration where all trials of the same speaker/item as a speech error trial or other missing trial were removed, 9.2% of the data was removed.

Speaker	Item	Verb relation	Object relation
DA001	1	New	New
DA001	5	Repeated	New
DA001	10	New	New
DA002	12	New	New
DA003	7	Repeated	New
DA003	7	Repeated	Old
DA004	11	Inference	New
DA005	10	New	New
DA005	10	Repeated	New
DA006	10	Inference	Old
DA007	3	New	New
DA010	4	Inference	New

Table J.1: Speech error trials excluded from phonetic analysis

J.2 Materials

For items 1-6, the intended inferencing relation linking the antecedent and target verb in the inference condition is entailment. For items 7-12, the intended relation is implicational bridging. The inferability scores for the included items as measured in Experiment 8 are: entailment/inferable: 6.56; entailment/discourse-new: 1.86; bridging/inferable: 5.49; bridging/discourse-new: 2.19.

Item 1

Opening context:	The kids rode the bus to the movies yesterday.
First clause:	
New verb, new object:	Eliza delayed Shirley, and
New verb, old object:	Samantha delayed Mary, and
Inference verb, new object:	Patty entertained Shirley, and
Inference verb, old object:	Asher entertained Mary, and
Repeated verb, new object:	Tiffany amused Shirley, and
Repeated verb, old object:	Jessica amused Mary, and
Target clause:	Will amused Mary.
Closing context:	They all agreed it was a great time.

Item 2

Opening context:	The trip to the corn maze was not so successful.
First clause:	
New verb, new object:	Carolyn guided Gary, and
New verb, old object:	Amanda guided Anna, and
Inference verb, new object:	Dylan aggravated Gary, and
Inference verb, old object:	James aggravated Anna, and
Repeated verb, new object:	Elena annoyed Gary, and
Repeated verb, old object:	Abigail annoyed Anna, and
Target clause:	Lynn annoyed Anna.
Closing context:	At least no one got lost!

Item 3

Opening context:	We had a great time at our college reunion.
First clause:	
New verb, new object:	Elijah rebuffed Eric, and
New verb, old object:	Andrea rebuffed Laura, and
Inference verb, new object:	Elizabeth hugged Eric, and
Inference verb, old object:	Veronica hugged Laura, and
Repeated verb, new object:	Oliver embraced Eric, and
Repeated verb, old object:	Christina embraced Laura, and

Target clause: Ron embraced Laura.
Closing context: We're looking forward to the next one.

Item 4

Opening context: My brother's birthday party got rowdy last week.
First clause:
 New verb, new object: Sebastian punished Justin, and
 New verb, old object: Gabriel punished Amy, and
 Inference verb, new object: Bradley astounded Justin, and
 Inference verb, old object: Ethan astounded Amy, and
 Repeated verb, new object: Julian surprised Justin, and
 Repeated verb, old object: Benjamin surprised Amy, and
Target clause: Nan surprised Amy.
Closing context: At least there's leftover cake to enjoy.

Item 5

Opening context: The huge line at the book signing went out the door.
First clause:
 New verb, new object: Adrian refused Ezra, and
 New verb, old object: Joshua refused Ryan, and
 Inference verb, new object: Susan idolized Ezra, and
 Inference verb, old object: Levi idolized Ryan, and
 Repeated verb, new object: Dominic admired Ezra, and
 Repeated verb, old object: Susan admired Ryan, and
Target clause: Neil admired Ryan.
Closing context: I don't know if I'll go to another one.

Item 6

Opening context: We had a tough time picking a job applicant.
First clause:
 New verb, new object: Ashley rejected Brenda, and
 New verb, old object: David rejected Emma, and
 Inference verb, new object: Rick recommended Brenda, and
 Inference verb, old object: Mike recommended Emma, and
 Repeated verb, new object: Zachary endorsed Brenda, and
 Repeated verb, old object: Mateo endorsed Emma, and
Target clause: Len endorsed Emma.
Closing context: We finally picked someone in the end, though.

Item 7

Opening context: Things were chaotic in my house the other day.
First clause:
 New verb, new object: Isabella dressed Felix, and
 New verb, old object: Alexander dressed Larry, and

Inference verb, new object:	Mohammed annoyed Felix, and
Inference verb, old object:	Elliot annoyed Larry, and
Repeated verb, new object:	Avery harassed Felix, and
Repeated verb, old object:	Everett harassed Larry, and
Target clause:	Ray harassed Larry.
Closing context:	Luckily, the kids all made it to school on time.

Item 8

Opening context:	The dance was just as crazy as we expected.
First clause:	
New verb, new object:	Matilda offended Griffin, and
New verb, old object:	Madeline offended Noah, and
Inference verb, new object:	Maximilian charmed Griffin, and
Inference verb, old object:	Angelina charmed Noah, and
Repeated verb, new object:	Marissa seduced Griffin, and
Repeated verb, old object:	Jocelyn seduced Noah, and
Target clause:	Al seduced Noah.
Closing context:	They should have more chaperones next time.

Item 9

Opening context:	We were all shocked by the drama at the hearing.
First clause:	
New verb, new object:	Brennan protected Cathy, and
New verb, old object:	Helen protected Allie, and
Inference verb, new object:	Kelsey misinformed Cathy, and
Inference verb, old object:	Betty misinformed Allie,
Repeated verb, new object:	Savannah deceived Cathy, and
Repeated verb, old object:	Natalie deceived Allie, and
Target clause:	Roy deceived Allie.
Closing context:	We're still waiting to hear the verdict.

Item 10

Opening context:	The police interrupted the school musical.
First clause:	
New verb, new object:	Ashley upstaged Alex, and
New verb, old object:	Vanesa upstaged Ollie, and
Inference verb, new object:	Olivia caught Alex, and
Inference verb, old object:	Valentina caught Ollie, and
Repeated verb, new object:	Serena exposed Alex, and
Repeated verb, old object:	Teresa exposed Ollie, and
Target clause:	May exposed Ollie.
Closing context:	We didn't even get to see the second act!

Item 11

Opening context: Bad behavior at the office reached new levels.
 First clause:
 New verb, new object: Maddison released Isaac, and
 New verb, old object: Miranda released Ernie, and
 Inference verb, new object: Sloan tolerated Isaac, and
 Inference verb, old object: Ruth tolerated Ernie, and
 Repeated verb, new object: Mallory condoned Isaac, and
 Repeated verb, old object: Rebecca condoned Ernie, and
 Target clause: Lee condoned Ernie.
 Closing context: Hopefully we can get back to normal soon.

Item 12

Opening context: The work mixer was surprisingly eventful.
 First clause:
 New verb, new object: Harrison involved Carol, and
 New verb, old object: Fiona involved Aaron, and
 Inference verb, new object: Betsy avoided Carol, and
 Inference verb, old object: Jacob avoided Aaron, and
 Repeated verb, new object: Jeremy disliked Carol, and
 Repeated verb, old object: Lucia disliked Aaron, and
 Target clause: Wayne disliked Aaron.
 Closing context: I'm glad we have the day off today.

J.3 Results and analysis for relativized duration

Recall that the analysis of duration data presented for Experiment 9 used raw duration values. One possible concern with this analysis is that raw duration values might depend on factors independent of the critical experimental manipulations, such as speaker-specific characteristics or changes in a speaker's speech rate over the course of the experiment. These concerns are partially alleviated by the fact that the experimental design was balanced within each participant, so every speaker produced each item in all six condition combinations. However, it is still prudent to verify that the results of the analysis of raw duration still hold when the duration values are relativized.

Figures J.1 and J.2 show the Experiment 9 results for the duration of each nucleus expressed as a ratio to the duration of the second-clause subject nucleus. Figure J.1 shows

the data with major speech error trials as well as other productions of error-containing items by the same speaker removed, as was done in the Experiment 9 analysis of raw duration. Figure J.2 shows the data with only the actual speech error trials removed.

Two factors led to the selection of this syllable as a standard for relativization. First, only the three words of the second clause had their interval boundaries manually corrected in the post-forced aligning TextGrid. This means that the duration values on this syllable are more likely to be correct than those on, for instance, the first-clause subject. Second, although some second-clause subjects carry a nuclear pitch accent and others a pre-nuclear accent, recall that no differences were found between subjects in different production conditions in any phonetic analyses or in the perception analysis from Experiment 10. This suggests that the duration of the second-clause subject is fairly robust across conditions and expressing the duration of the other syllables relative to this duration will faithfully represent true differences between the conditions on those syllables.

The results of the statistical analysis for relative duration with error trials and items with the same item/speaker combination removed are as follows. **O1:** marginal effect of object status ($p < .1$) for entailment items. Verb status and interaction n.s. (p 's $> .4$). No significant effect of verb status, object status, or interaction for bridging items (p 's $> .2$).

O2: marginal effect of object status ($p < .1$) for entailment items. Verb status and interaction n.s. (p 's $> .4$). No significant effects for bridging items (p 's $> .2$).

V1: no significant effects of verb status, object status, or interaction for either entailment or bridging items (p 's $> .2$).

V2, entailment items: significant effect of verb status ($p < .01$). Paired comparisons with estimated marginal means: new-repeated significant ($p < .01$), new-inference n.s. ($p > .1$), inference-repeated n.s. ($p > .2$). No significant effect of object status or interaction (p 's $> .3$).

V2, bridging items: significant interaction of verb status and object status ($p > .05$), significant effects of verb status and object status (p 's $< .001$). Paired comparisons with es-

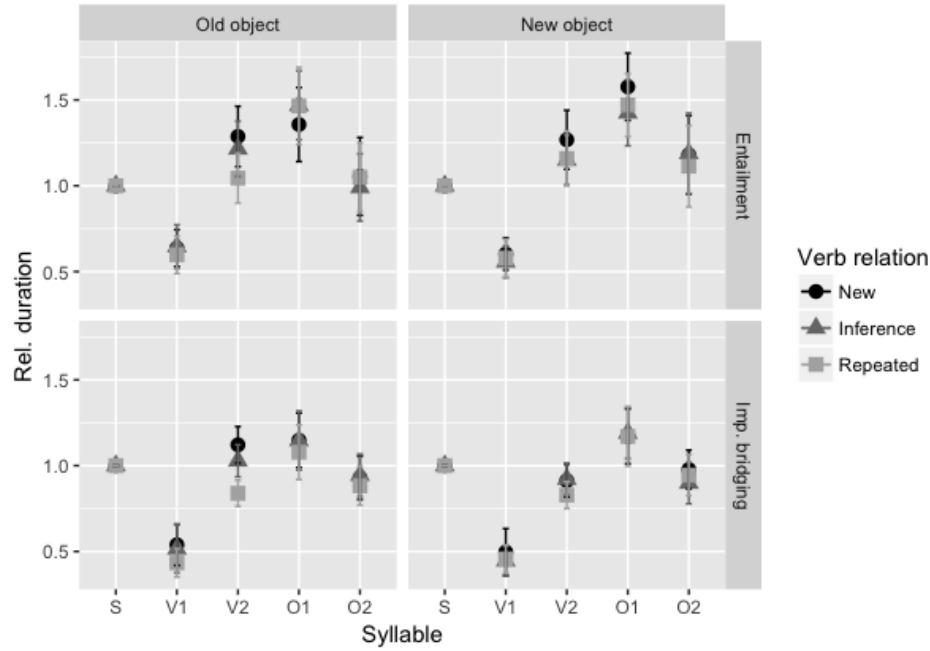


Figure J.1: Experiment 9 results for duration relative to second-clause subject with all data matching item/speaker combination of an error trial removed. Horizontal position: Syllable within critical clause. Horizontal split: Object status. Vertical split: Inferencing relation (items 1-6 versus 7-12). Error bars: 95% confidence interval.

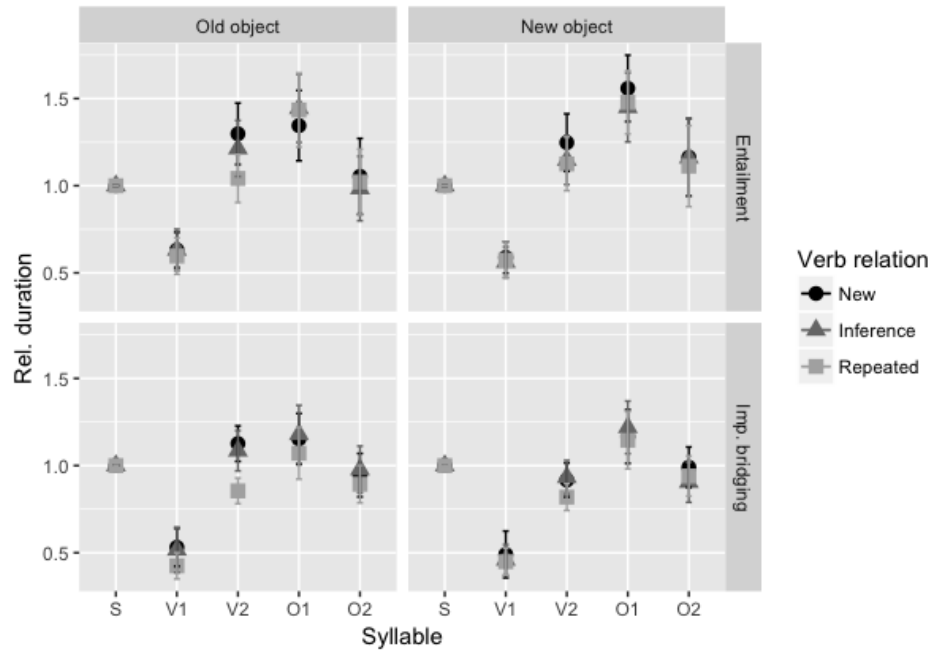


Figure J.2: Experiment 9 results for duration relative to second-clause subject with all non-error data. Horizontal position: Syllable within critical clause. Horizontal split: Object status. Vertical split: Inferencing relation (items 1-6 versus 7-12). Error bars: 95% confidence interval.

estimated marginal means within old-object conditions: new-repeated significant ($p < .001$), new-inference n.s. ($p > .1$), inference-repeated significant ($p < .01$). Within new-object conditions: no comparisons significant (p 's $> .1$).

Thus, the analysis of the relativized duration data substantially reduces the significance values of several of the critical comparisons compared to the raw duration analysis. In particular, the accent contrast between new and old objects is marginal to not significant for relative duration, whereas it was highly significant for raw duration. In addition, the critical verb finding (similar values for new and inference to the exclusion of repeated on old-object V2) was not detected for entailment items, although a numeric trend in favor of this effect is apparent from the plots. However, the same trend was observed for the bridging items as in the raw duration data.

Thus, the results of the relative duration analysis are somewhat mixed. The critical interaction for verbs and, among old-object verbs, theoretically interesting division between new/inference verbs and repeated verbs was not detected for entailment items. This is likely driven by the higher values on new verbs with a new object, which prevented the verb/object interaction from attaining significance. The high values for inference verbs with an old object and lower values for inference verbs with a new object led to a conclusion that the inference values were, across the board, intermediate between those for new and repeated verbs on the entailment items.

In addition, for the paradigm check, the emphasis difference between new and old objects, the critical effect was not detected in the relative duration analysis. On the surface, these two findings suggest that the critical results of the raw duration analysis should be taken with a grain of salt, as several otherwise reliable effects disappear when the data are relativized. However, the fact that the object effect was reliably detected in the analysis of z -intensity, z - f_0 , raw duration, and the Experiment 10 perception results suggests that the relative duration analysis is not sensitive enough to detect significant differences in the production data (e.g., because the range of possible data values is too compressed).

Furthermore, the characteristic verb results observed in all of the other analyses is still apparent on V2 for the bridging items in the relative duration analysis. This suggests that at least for the bridging items, the findings reported in the main body of the dissertation for Experiment 9 are robust and the broad conclusions discussed there are valid. However, it may be worth exploring other strategies for normalizing the Experiment 9 duration results in the future to further ensure the validity of the associated conclusions.

APPENDIX K

EXPERIMENT 10: PERCEPTION OF EMPHASIS IN PRODUCTION DATA

K.1 Task and layout

The sentence you will hear is:

Nan surprised Amy.

Pay attention to the word:

Amy

Press any key to hear the sentence.

In this sentence, was the word *Amy* emphasized or not emphasized?

1. [emphasized](#)
2. [not emphasized](#)

Figure K.1: Sample screens from Experiment 10. The sound file (e.g., *Nan surprised Amy*) played automatically when the second screen appeared.

K.2 Participant statistics and data loss

Participants:	201
Gender:	70 female, 131 male
Mean age:	33.7 years
Non-native speakers excluded:	10 (5.0%)
Attention criterion 1:	Participant’s mean reaction time $\leq$ 500 ms
Inattentive participants excluded:	8 (4.0%)
Attention criterion 2:	At least 50% concordance with intended emphasis rating on fillers
Additional inattentive participants excluded:	8 additional (4.0%)
Average completion time (included participants):	6 min, 32 sec
Payment:	USD 1.00

Data using recordings from the production items shown in Table J.1 was excluded from analysis due to major speech errors identified during manual correction of the f_0 values. This removed 1.6% of the data that remained after the participant-level exclusions described above. For the trial that was excluded from Experiment 9 due to a labeling error, the clause recording was manually clipped and is included in Experiment 10.

K.3 Materials

The materials were clipped recordings of the second *SVO* clauses taken from the production study, Experiment 9. The item numbers and condition labels are the same as for that experiment and can be seen in Appendix J.

APPENDIX L

EXPERIMENT 11: OUT-OF-THE-BLUE DEACCENTING PERCEPTION

L.1 Task and layout

The sentence you will hear is:

Sebastian punished Justin, and Nan surprised Amy.

Press any key to hear the sentence.

On a scale from 1 to 7, where 1 is the least natural and 7 is the most natural, how natural did you find the "melody" or "tune" of this sentence?

Less natural *More natural*

Figure L.1: Sample screens from Experiment 11. The critical sentence recording played automatically when the second screen was loaded.

L.2 Participant statistics and data loss

Participants:	144
Gender:	69 female, 75 male
Mean age:	36.6 years
Non-native speakers excluded:	1 (0.7%)
Attention criterion:	Participant's mean reaction time $\leq$ 1000 ms
Additional inattentive participants excluded:	5 (3.5%)
Average completion time (included participants):	8 min, 1 sec
Payment:	USD 1.00

L.3 Additional recording stimuli

The stimuli for Experiments 11 through 13 were recorded using the same paradigm as was used for Experiment 9. Each of the two speakers recorded the additional stimuli in three sessions. Items 13 through 24 were recorded in the first additional session, items 25 through 36 in the second session, and 37 through 48 in the fourth session.

Item 13

Opening context:	The lottery winners and losers were announced.
First clause:	
New verb, new object:	Mark applauded Anthony, and
New verb, old object:	Ben applauded Erica, and
Inference verb, new object:	Mary consoled Anthony, and
Inference verb, old object:	Harry consoled Erica, and
Repeated verb, new object:	Sam comforted Anthony, and
Repeated verb, old object:	Bill comforted Erica, and
Target clause:	Ted comforted Erica.
Closing context:	A hundred-million-dollar prize awaited the winner.

Item 14

Opening context:	The big table was completely set for poker.
First clause:	
New verb, new object:	Carlos affirmed Molly, and
New verb, old object:	Marcos affirmed Sally, and
Inference verb, new object:	Abigail scammed Molly, and
Inference verb, old object:	Kennedy scammed Sally, and
Repeated verb, new object:	Alfie cheated Molly, and
Repeated verb, old object:	Maxwell cheated Sally, and
Target clause:	Reagan cheated Sally.
Closing context:	I don't think they were still friends after that night.

Item 15*

Opening context:	The class had P.E. during second period.
First clause:	
New verb, new object:	Benjamin picked Roger, and
New verb, old object:	Shirley-Ann picked Martha, and
Inference verb, new object:	Elijah floored Susie, and
Inference verb, old object:	Martina floored Martha, and
Repeated verb, new object:	Sam astonished Millie, and
Repeated verb, old object:	Wayne astonished Martha, and
Target clause:	Greg astonished Martha.

Closing context: The teacher enjoyed watching the students go wild in a basketball game.

Item 16

Opening context: The plumbers entered the next bathroom with their tools.

First clause:

New verb, new object: Chad embarrassed Karen, and

New verb, old object: Bill embarrassed Linda, and

Inference verb, new object: Oliver billed Karen, and

Inference verb, old object: Gabriel billed Linda, and

Repeated verb, new object: Anthony charged Karen, and

Repeated verb, old object: Sebastian charged Linda, and

Target clause: Joseph charged Linda.

Closing context: There should have been a plumbing inspection months ago.

Item 17

Opening context: The kids threw a ball and smashed the vase on the floor.

First clause:

New verb, new object: Christopher feared Pauline, and

New verb, old object: Anthony feared Amy, and

Inference verb, new object: Samuel scolded Pauline, and

Inference verb, old object: Jess scolded Amy, and

Repeated verb, new object: Don admonished Pauline, and

Repeated verb, old object: Sal admonished Amy, and

Target clause: Chris admonished Amy.

Closing context: They should never play baseball inside again!

Item 18

Opening context: The marching band started early morning practice.

First clause:

New verb, new object: Emma impressed Leslie, and

New verb, old object: Joanne impressed Andy, and

Inference verb, new object: Amelia woke Leslie, and

Inference verb, old object: Eliza woke Andy, and

Repeated verb, new object: Alexis roused Leslie, and

Repeated verb, old object: Jeremy roused Andy, and

Target clause: Cecil roused Andy.

Closing context: Maybe they should start at 8 in the morning instead of 5.

Item 19

Opening context: The guests moved over to the wedding reception.

First clause:
 New verb, new object: Demetrius upset Elizabeth, and
 New verb, old object: Alexander upset Olivia, and
 Inference verb, new object: Paul congratulated Elizabeth, and
 Inference verb, old object: Karl congratulated Olivia, and
 Repeated verb, new object: Justin complimented Elizabeth, and
 Repeated verb, old object: Frances complimented Olivia, and
 Target clause: George complimented Olivia.
 Closing context: The newlyweds were excited to greet people after the ceremony.

Item 20

Opening context: The kids started a kickball game during recess.
 First clause:
 New verb, new object: Jonathan accepted Jane, and
 New verb, old object: Marissa accepted Roy, and
 Inference verb, new object: Veronica bullied Jane, and
 Inference verb, old object: Alexandra bullied Roy, and
 Repeated verb, new object: Anne intimidated Jane, and
 Repeated verb, old object: Ruth intimidated Roy, and
 Target clause: Kendall intimidated Roy.
 Closing context: Schoolyard scuffles and games go hand in hand, it seems.

Item 21

Opening context: We just had a family therapy appointment.
 First clause:
 New verb, new object: Dahlia echoed Lily, and
 New verb, old object: Zachary echoed Anna, and
 Inference verb, new object: Benny frustrated Lily, and
 Inference verb, old object: Matthew frustrated Anna, and
 Repeated verb, new object: Elise discouraged Lily, and
 Repeated verb, old object: Chloe discouraged Anna, and
 Target clause: Nate discouraged Anna.
 Closing context: I think another appointment would be good for the family.

Item 22

Opening context: We visited my family in Arizona.
 First clause:
 New verb, new object: Maggie threatened Will, and
 New verb, old object: Rosa threatened Eve, and
 Inference verb, new object: Gaia indulged Will, and
 Inference verb, old object: Jackie indulged Eve, and
 Repeated verb, new object: Clara spoiled Will, and

Repeated verb, old object: Mickey spoiled Eve, and
Target clause: Walter spoiled Eve.
Closing context: We're not going back to the Southwest for a while.

Item 23

Opening context: The parents brought their children for a fun playdate.
First clause:
New verb, new object: Rosalie drew Bella, and
New verb, old object: Samantha drew Phoebe, and
Inference verb, new object: Pete reminded Bella, and
Inference verb, old object: Mark reminded Phoebe, and
Repeated verb, new object: Marina nagged Bella, and
Repeated verb, old object: Theodore nagged Phoebe, and
Target clause: Mira nagged Phoebe.
Closing context: The parents scheduled another one for next Saturday.

Item 24

Opening context: The science clubs held joint meetings for everyone.
First clause:
New verb, new object: Eli included Ethan, and
New verb, old object: Adelle included Mason, and
Inference verb, new object: Logan refuted Ethan, and
Inference verb, old object: Ali refuted Mason, and
Repeated verb, new object: Lucas corrected Ethan, and
Repeated verb, old object: Liam corrected Mason, and
Target clause: Daisy corrected Mason.
Closing context: I think clear communication would be better for everyone in the future.

Item 25

Opening context: The competitive game show became more heated.
First clause:
New verb, new object: Scarlett helped Bill, and
New verb, old object: Minerva helped Doug, and
Inference verb, new object: Morgan despised Bill, and
Inference verb, old object: Marcus despised Doug, and
Repeated verb, new object: Seamus hated Bill, and
Repeated verb, old object: Thomas hated Doug, and
Target clause: Mary hated Doug.
Closing context: It was an exciting episode for the viewers.

Item 26

Opening context: The players on the sidelines viewed the World Cup match.

First clause:
 New verb, new object: Ben replaced Asher, and
 New verb, old object: Dan replaced Abby, and
 Inference verb, new object: Will watched Asher, and
 Inference verb, old object: Max watched Abby, and
 Repeated verb, new object: Donald saw Asher, and
 Repeated verb, old object: Angus saw Abby, and
 Target clause: Leo saw Abby.
 Closing context: If only the coach put the benchwarmers on the field.

Item 27

Opening context: The Halloween costume party was at my house.
 First clause:
 New verb, new object: Darren pleased Johnny, and
 New verb, old object: Maggie pleased Jeffrey, and
 Inference verb, new object: Zach terrified Johnny, and
 Inference verb, old object: Mark terrified Jeffrey, and
 Repeated verb, new object: Donny scared Johnny, and
 Repeated verb, old object: Allan scared Jeffrey, and
 Target clause: Frank scared Jeffrey.
 Closing context: I want to host it again next year.

Item 28

Opening context: We hired bad camp counselors for the summer.
 First clause:
 New verb, new object: Harrison sheltered Steven, and
 New verb, old object: Raphael sheltered Mabel, and
 Inference verb, new object: Zander perverted Steven, and
 Inference verb, old object: Norman perverted Mabel, and
 Repeated verb, new object: Tammy corrupted Steven, and
 Repeated verb, old object: Hubert corrupted Mabel, and
 Target clause: Jordan corrupted Mabel.
 Closing context: We might never hire from high schools again.

Item 29

Opening context: Most of the kitchen workers were catching the flu.
 First clause:
 New verb, new object: Angelica served Mac, and
 New verb, old object: Jeremiah served Rex, and
 Inference verb, new object: Katarina tainted Mac, and
 Inference verb, old object: Elizabeth tainted Rex, and
 Repeated verb, new object: Don contaminated Mac, and
 Repeated verb, old object: Moe contaminated Rex, and
 Target clause: James contaminated Rex.

Closing context: The entire staff should take the week off from work.

Item 30

Opening context: A large network of spies entered the royal court.

First clause:

New verb, new object: Rebecca fed Mimi, and
New verb, old object: Benjamin fed Arthur, and
Inference verb, new object: Al sabotaged Mimi, and
Inference verb, old object: Abe sabotaged Arthur, and
Repeated verb, new object: Chris undermined Mimi, and
Repeated verb, old object: Shane undermined Arthur, and

Target clause: Mark undermined Arthur.

Closing context: The intrigue was too much for the king to handle.

Item 31

Opening context: After the scandal, reputations were in doubt.

First clause:

New verb, new object: Oliver duped Emma, and
New verb, old object: Elijah duped Daniel, and
Inference verb, new object: Henry supported Emma, and
Inference verb, old object: Lukas supported Daniel, and
Repeated verb, new object: Simone defended Emma, and
Repeated verb, old object: Aaron defended Daniel, and

Target clause: Lee defended Daniel.

Closing context: I don't know if we'll be able to clear everyone's names.

Item 32

Opening context: I don't know if we'll be able to clear everyone's names.

First clause:

New verb, new object: Josh instructed Vicky, and
New verb, old object: Liz instructed Owen, and
Inference verb, new object: Carter ignored Vicky, and
Inference verb, old object: Aidan ignored Owen, and
Repeated verb, new object: Anne excluded Vicky, and
Repeated verb, old object: Joe excluded Owen, and

Target clause: Sebastian excluded Owen.

Closing context: It was the first time 4th and 10th graders had studied together.

Item 33

Opening context: The day for student presentations had arrived.

First clause:

New verb, new object: Francis encouraged Alice, and
New verb, old object: Eva encouraged Isaac, and

Inference verb, new object:	Faye interrupted Alice, and
Inference verb, old object:	Jade interrupted Isaac, and
Repeated verb, new object:	Gabriel heckled Alice, and
Repeated verb, old object:	Vivian heckled Isaac, and
Target clause:	Jeb heckled Isaac.
Closing context:	The audience was very lively!

Item 34

Opening context:	The therapist's office opened for a new day.
First clause:	
New verb, new object:	Lynne fetched Edward, and
New verb, old object:	Joshua fetched Dylan, and
Inference verb, new object:	Belle placated Edward, and
Inference verb, old object:	Rob placated Dylan, and
Repeated verb, new object:	Helena calmed Edward, and
Repeated verb, old object:	Adela calmed Dylan, and
Target clause:	Ron calmed Dylan.
Closing context:	It was a busy day for the therapists, but every appointment happened.

Item 35

Opening context:	The police found the criminal gang in a bar.
First clause:	
New verb, new object:	Evelyn ignored Sophie, and
New verb, old object:	Emily ignored Anders, and
Inference verb, new object:	Isabelle detained Sophie, and
Inference verb, old object:	Abigail detained Anders, and
Repeated verb, new object:	Harper arrested Sophie, and
Repeated verb, old object:	Ingrid arrested Anders, and
Target clause:	Gene arrested Anders.
Closing context:	All of the lawbreakers were apprehended.

Item 36

Opening context:	The lawyers stood up to question the witnesses.
First clause:	
New verb, new object:	Kennedy grilled Lucy, and
New verb, old object:	Adrienne grilled Zoey, and
Inference verb, new object:	Sal fascinated Lucy, and
Inference verb, old object:	Matt fascinated Zoey, and
Repeated verb, new object:	Nicholas dazzled Lucy, and
Repeated verb, old object:	Isaiah dazzled Zoey, and
Target clause:	Caleb dazzled Zoey.
Closing context:	It was an impressive trial to watch.

Item 37

Opening context: My kids played war with our neighbors' in the backyard.
First clause:
 New verb, new object: Julius strengthened Maya, and
 New verb, old object: Emily strengthened Mia, and
 Inference verb, new object: Alana captured Maya, and
 Inference verb, old object: Delilah captured Mia, and
 Repeated verb, new object: Ben apprehended Maya, and
 Repeated verb, old object: Tim apprehended Mia, and
Target clause: Sid apprehended Mia.
Closing context: All of them ended up with bruises and scrapes afterwards.

Item 38

Opening context: The troops broke into the prisoner of war camp.
First clause:
 New verb, new object: Matthew scared Alex, and
 New verb, old object: Alli scared Ivy, and
 Inference verb, new object: Smith released Alex, and
 Inference verb, old object: Sue released Ivy, and
 Repeated verb, new object: Eric freed Alex, and
 Repeated verb, old object: Hunter freed Ivy, and
Target clause: Terra freed Ivy.
Closing context: Many of the soldiers cried after gaining their freedom.

Item 39

Opening context: Multiple contestants entered the boxing ring.
First clause:
 New verb, new object: Steph supported Jack, and
 New verb, old object: Ray supported Jake, and
 Inference verb, new object: Jerry dodged Jack, and
 Inference verb, old object: Christine dodged Jake, and
 Repeated verb, new object: Ted evaded Jack, and
 Repeated verb, old object: Scott evaded Jake, and
Target clause: Mike evaded Jake.
Closing context: This slugfest would go down in history.

Item 40

Opening context: At the theater, a sad part of the movie came.
First clause:
 New verb, new object: Katherine surprised Helen, and
 New verb, old object: Melissa surprised Ryan, and
 Inference verb, new object: Hannah pacified Helen, and
 Inference verb, old object: Laura pacified Ryan, and

Repeated verb, new object: Peter quieted Helen, and
Repeated verb, old object: Harold quieted Ryan, and
Target clause: Danny quieted Ryan.
Closing context: The audience reacted to the scene pretty strongly!

Item 41

Opening context: The New York team was losing to Los Angeles.
First clause:
New verb, new object: Harry taunted Ronald, and
New verb, old object: Kevin taunted Kenneth, and
Inference verb, new object: Gary soothed Ronald, and
Inference verb, old object: Maxwell soothed Kenneth, and
Repeated verb, new object: Camille calmed Ronald, and
Repeated verb, old object: Nancy calmed Kenneth, and
Target clause: Dillon calmed Kenneth.
Closing context: The scores were not very close at all.

Item 42

Opening context: The United Nations diplomats met.
First clause:
New verb, new object: Edward contacted Michelle, and
New verb, old object: Samson contacted Halley, and
Inference verb, new object: Jason detested Michelle, and
Inference verb, old object: Ashley detested Halley, and
Repeated verb, new object: Angela hated Michelle, and
Repeated verb, old object: Pamela hated Halley, and
Target clause: Virginia hated Halley.
Closing context: Thankfully they were able to agree on something.

Item 43

Opening context: The superhero entered the bad guys' hideout.
First clause:
New verb, new object: Tyler telephoned Tony, and
New verb, old object: Cooper telephoned Gavin, and
Inference verb, new object: Zachary subdued Tony, and
Inference verb, old object: Arianna subdued Gavin, and
Repeated verb, new object: Chase overpowered Tony, and
Repeated verb, old object: Jace overpowered Gavin, and
Target clause: Adam overpowered Gavin.
Closing context: The world was safe, for today at least.

Item 44**

Opening context: The secret agents infiltrated the prison.

First clause:
 New verb, new object: Nathaniel telephoned Roman, and
 New verb, old object: Leonardo rescued Austin, and
 Inference verb, new object: Elias influenced Roman, and
 Inference verb, old object: Dominic influenced Austin, and
 Repeated verb, new object: Cole manipulated Roman, and
 Repeated verb, old object: Cal manipulated Austin, and
 Target clause: Nolan manipulated Austin.
 Closing context: The agency would reward the agents for accomplishing their mission.

Item 45

Opening context: The Hollywood bigshots held exclusive mixers.
 First clause:
 New verb, new object: Joel represented Oscar, and
 New verb, old object: Finn represented Sarah, and
 Inference verb, new object: Elliot adored Oscar, and
 Inference verb, old object: Melanie adored Sarah, and
 Repeated verb, new object: Alejandro loved Oscar, and
 Repeated verb, old object: Emmanuel loved Sarah, and
 Target clause: Victor loved Sarah.
 Closing context: One of those parties should be made into a movie!

Item 46

Opening context: The actors gathered after the last performance.
 First clause:
 New verb, new object: Faith concerned Ashe, and
 New verb, old object: Clare concerned Gale, and
 Inference verb, new object: Riley praised Ashe, and
 Inference verb, old object: Julie praised Gale, and
 Repeated verb, new object: Val flattered Ashe, and
 Repeated verb, old object: Graham flattered Gale, and
 Target clause: Cora flattered Gale.
 Closing context: The actors thought they had done an excellent job.

Item 47

Opening context: The magicians brought their show into town Tuesday.
 First clause:
 New verb, new object: Grant described Luna, and
 New verb, old object: Quinn described Gia, and
 Inference verb, new object: Hayden fooled Luna, and
 Inference verb, old object: Ruby fooled Gia, and
 Repeated verb, new object: Kylie conned Luna, and
 Repeated verb, old object: Brian conned Gia, and

Target clause:	Peyton conned Gia.
Closing context:	The mayor won't let the magic show back into town ever again.

Item 48

Opening context:	One of my old friends hosted a birthday party.
First clause:	
New verb, new object:	Jasper controlled Eden, and
New verb, old object:	August controlled Jesse, and
Inference verb, new object:	Athena approached Eden, and
Inference verb, old object:	Maximus approached Jesse, and
Repeated verb, new object:	Ariel greeted Eden, and
Repeated verb, old object:	Abraham greeted Jesse, and
Target clause:	Dean greeted Jesse.
Closing context:	It was nice to see everyone again after so long.

*Due to experimenter error, item 15 featured three different names as objects in the first clause of new-object conditions, whereas the other items featured the same name in this position. However, this did not affect the discourse status of any of the constituents in the second clause, so the item was retained for the perception studies.

**Due to experimenter error, item 44 featured different first-clause verbs in the new-object/new-verb versus old-object/new-verb conditions. However, this did not affect the discourse status of any of the constituents in the second clause, so the item was retained for the perception studies.

L.4 Critical stimuli

As described in section 7.1, the critical stimuli were constructed by cross splicing a new-verb or repeated-verb second clause of a particular object status to after recorded first clauses of all three verb status levels and the same object status. The resulting sentences were string-identical to the critical sentences shown in Appendix J and the re-recorded stimuli immediately above. The second-clause object was canonically realized with accent if new and without accent if repeated, while the second-clause verb instantiated a 3 (discourse status) x 2 (accented/deaccented) design.

L.5 Fillers

The experiment included 10 fillers. 5 were designed to have felicitous prosody, and 5 were designed to be infelicitous. The fillers are listed below. The infelicitous fillers are accompanied by a brief description of the key prosody leading to infelicitousness.

Item 49	At the zoo, we saw rhinos, apes, and wolves.	Felicitous
Item 50	Buy me an apple.	Felicitous
Item 51	Have you finished your homework yet?	Felicitous
Item 52	Who did you see at the movies?	Felicitous
Item 53	If you have any problems, just contact us.	Felicitous
Item 54	Did you see the rainbow yesterday? (Falling pitch on <i>yesterday</i>)	Infelicitous
Item 55	What did you decide to buy? (Falling pitch on <i>did</i> , rising pitch on <i>buy</i>)	Infelicitous
Item 56	Go clean your room. (Rising pitch on <i>room</i>)	Infelicitous
Item 57	Yesterday I bought eggs, milk, and cheese. (Falling pitch on <i>eggs</i> and <i>milk</i> , rising pitch on <i>cheese</i>)	Infelicitous
Item 58	What a beautiful voice. (Rising pitch on <i>voice</i>)	Infelicitous

APPENDIX M

EXPERIMENT 12: DEACCENTING PERCEPTION IN CONTEXT, I

M.1 Task and layout

Here is a scenario:

As they did every year, the teachers worried about how the students would interact with each other on the first day of high school.

Press any key to hear a speaker talk about the scenario.

On a scale from 1 to 7, where 1 is the least natural and 7 is the most natural, how natural did you find the "melody" or "tune" of this sentence?

Less natural *More natural*

Figure M.1: Sample screens from Experiment 12. The critical sentence recording played automatically when the second screen was loaded.

M.2 Participant statistics and data loss

Participants:	144
Gender:	53 female, 91 male
Mean age:	33.5 years
Non-native speakers excluded:	2 (1.4%)
Attention criterion:	Participant's mean reaction time $\leq$ 1000 ms
Additional inattentive participants excluded:	7 (4.9%)
Average completion time (included participants):	8 min, 28 sec
Payment:	USD 1.00

M.3 Critical items

With item numbers corresponding to the stimuli shown in Appendices J and L, items 11, 12, 14, 23, 24, 26, 28, 29, 30, 37, 45, and 47 were not included in the experiment, as discussed

in section 7.2. The context sentences that were added before the critical recording in the remaining items are as follows.

- Item 1** As the kids gathered together to put on a show, only some were in a playful mood.
- Item 2** The staff members generally got along with the new hires, although there were some personality clashes.
- Item 3** The high school reunion was very eventful, with many people seeing each other for the first time in ten years.
- Item 4** After the disastrous swim meet, the team wasn't sure what treatment to expect from their coaches.
- Item 5** As the annual meeting drew to a close, the writers on the Lifetime Achievement Award committee all voiced their opinions on the different candidates.
- Item 6** Since the stakes were so high, the hiring committee spent a long time discussing the merits of each candidate.
- Item 7** The kids were driving each other crazy yesterday, although some were in a helpful mood.
- Item 8** Some people were successful at the speed dating event, and some weren't.
- Item 9** Since the young girls had been orphaned years ago, they had no choice but to trust their legal guardians.
- Item 10** The spies were always trying to one-up each other, and ending another spy's career was considered the ultimate prize.
- Item 13** The team lost the game, but they were able to focus on the positive elements.
- Item 15** Everybody's future was up in the air during the coaches' scouting visit, and nobody knew what to expect.
- Item 16** The workers at the trade show were supposed to be giving away free samples from their companies, but they were surprisingly hostile toward women.
- Item 17** Although there had been several workshops, the hostile environment at our office didn't do much for morale.
- Item 18** The students used to be indifferent toward volunteering, but felt differently after meeting with the famous activists.
- Item 19** After the performers finished the recital, some of the audience members shared their reactions.
- Item 20** As they did every year, the teachers worried about how the students would interact with each other on the first day of high school.
- Item 21** The directors thought the performance went very badly, but the actors defended themselves afterwards.
- Item 22** A lot of the kids at the family reunion were being naughty, and it was interesting to observe everyone's parenting style.
- Item 25** Relations among the childhood friends were all over the place after the big fight that broke out yesterday.

- Item 27** Just like every year, we had a great time exchanging gag gifts on April Fool's Day.
- Item 31** Like every year, there were a lot of clever maneuvers during the legislative session.
- Item 32** The camp counselors who led the session were disappointed that the kids weren't getting along during archery lessons.
- Item 33** Audience participation was encouraged at the experimental theater, and last night's audience had a lot to say during the show.
- Item 34** To their credit, everyone jumped right into action after the attack.
- Item 35** Tensions were high between the suspects after the officers arrived.
- Item 36** As everyone expected, the performances at the mock trial were very impressive.
- Item 38** After the fight, different teachers had their own approaches to disciplinary problems.
- Item 39** All the training paid off, and our side displayed a lot of effective tactics during the flag football game.
- Item 40** As they usually did, the protestors at the town hall meeting expected resistance to their complaints.
- Item 41** The teammates worked to support each other after the close loss, but not everybody was very sportsmanlike.
- Item 42** Everyone had a different reaction when the shocking video was first released.
- Item 43** The officers decided to split up, with some approaching the suspects and others calling for backup.
- Item 44** At the most recent meeting of the political committee, the younger members realized how easy it is to fall into a trap.
- Item 46** Just like every year, the teachers spent all of parent conference night delivering good or bad news about the students.
- Item 48** The security guards did their best to manage the crowds during the celebrity meet-and-greet.

M.4 Fillers

The experiment used the same filler recordings shown in Appendix L. To match the critical task, the fillers were preceded with the following context sentences.

- Item 49** Since the weather was so beautiful, we decided to do something outdoors.
- Item 50** Are you going to the corner store?
- Item 51** I know you want to go to the party with your friends.
- Item 52** I'm glad to see you made it home safely.
- Item 53** The service team is always here to support you.
- Item 54** We were all amazed by how quickly the weather changed.

- Item 55** You're back from the store already?
- Item 56** Your grandparents will be here any minute.
- Item 57** I should be able to prepare my favorite recipe.
- Item 58** We went to an excellent vocal recital the other day.

APPENDIX N

EXPERIMENT 13: DEACCENTING PERCEPTION IN CONTEXT, II

N.1 Task and layout

Imagine you just heard someone say:

Like every year, there were a lot of clever maneuvers during the legislative session.

Next, press any key to hear the same person say:

Henry supported Emma, and Lee defended Daniel.

On a scale from 1 to 7, where 1 is the least natural and 7 is the most natural, how natural did you find the "melody" or "tune" of this sentence?

Less natural *More natural*

Figure N.1: Sample screens from Experiment 13. The critical sentence recording played automatically when the second screen was loaded.

N.2 Participant statistics and data loss

Participants:	145
Gender:	53 female, 92 male
Mean age:	35.6 years
Non-native speakers excluded:	1 (0.7%)
Attention criterion:	Participant's mean reaction time $\leq$ 1000 ms
Additional inattentive participants excluded:	10 (6.9%)
Average completion time (included participants):	9 min, 25 sec
Payment:	USD 1.00

N.3 Materials

The context sentences and critical sentences/recordings were identical to those used in Experiment 12 and outlined in Appendix M. The materials were displayed differently in Experiment 13 as described in section 7.2.

APPENDIX O

EXPERIMENT 14: DEACCENTING PERCEPTION WITH
PRESUPPOSITION TRIGGERS

O.1 Task and layout

The sentence you will hear is:

Chloe discouraged Anna, and Nate discouraged Anna, too.

Press any key to hear the sentence.

On a scale from 1 to 7, where 1 is the least natural and 7 is the most natural, how natural did you find the "melody" or "tune" of this sentence?

Less natural *More natural*

Figure O.1: Sample screens from Experiment 14. The critical sentence recording played automatically when the second screen was loaded.

O.2 Participant statistics and data loss

Participants:	140
Gender:	46 female, 94 male
Mean age:	37.0 years
Non-native speakers excluded:	0 (0.0%)
Attention criterion:	Participant's mean reaction time $\leq$ 1000 ms
Additional inattentive participants excluded:	23 (16.4%)
Average completion time (included participants):	8 min, 44 sec
Payment:	USD 1.00

O.3 Materials

The materials were recorded using the same paradigm as Experiment 9 and the re-recording before Experiment 11, but with a comma and *too* added to the end of each critical sentence. Only the female speaker participated, and the recordings were made remotely rather than in a sound booth. As before, the stimuli were recorded in four sessions, covering items 1-12, 13-24, 25-36, and 37-48, respectively. The critical items are identical to the design stimuli outlined for Experiment 11, but with *too* at the end of the sentence. The filler items are the same as Experiment 11, but were re-recorded by the female speaker, who imitated the prosody of the previous recordings.

REFERENCES

- Abrusán, M. (2014). On the focus-sensitive presupposition triggers too, again, also, even. *Proceedings of Sinn und Bedeutung 18*, pages 6–23.
- Adams, C. and Munro, R. R. (1978). In search of the acoustic correlates of stress: Fundamental frequency, amplitude, and duration in the connected utterances of some native and non-native speakers. *Phonetica*, 35:125–156.
- Aelbrecht, L. (2010). *The Syntactic Licensing of Ellipsis*. John Benjamins, Amsterdam.
- Allopenna, P. D., Magnuson, J. S., and Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye-movements: Evidence for continuous mapping models. *Journal of Memory and Language*, 38:419–439.
- Ariel, M. (1988). Referring and accessibility. *Journal of Linguistics*, 24:67–87.
- Ariel, M. (1990). *Accessing Antecedents*. Routledge, London.
- Arregui, A., Clifton Jr., C., Frazier, L., and Moulton, K. (2006). Processing elided verb phrases with flawed antecedents: The recycling hypothesis. *Journal of Memory and Language*, 55:232–246.
- Bard, E. and Aylett, M. (1999). The dissociation of de-accenting, givenness, and syntactic role in spontaneous speech. *Proceedings of the the Fourteenth International Conference of Phonetic Sciences*, pages 1753–1756.
- Baumann, S. and Riester, A. (2012). Referential and lexical givenness: Semantic, prosodic and cognitive aspects. In Elordieta, G. and Prieto, P., editors, *Prosody and Meaning*, pages 1–34. Mouton de Gruyter, Berlin.
- Beaver, D. and Clark, B. C. (2008). *Sense and Sensitivity: How Focus Determines Meaning*. Wiley-Blackwell, Malden, MA.
- Beckman, M. E. (1986). *Stress and Non-stress Accent*. Foris Publications, Dordrecht.
- Beckman, M. E. and Edwards, J. (1994). Articulatory evidence for differentiating stress categories. *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology III*, pages 7–33.
- Bergen, L. and Goodman, N. D. (2015). The strategic use of noise in pragmatic reasoning. *Topics in Cognitive Science*, 7:336–350.
- Berman, A. and Szamosi, M. (1972). Observations on sentential stress. *Language*, 48:304–325.
- Bever, T. G. (1970). The cognitive basis for linguistic structures. In Hayes, J., editor, *Cognition and the Development of Language*, pages 279–362. Wiley, New York.
- Boersma, P. and Weenink, D. (2013). Praat, a system for doing phonetics by computer.

- Bolinger, D. L. (1958). A theory of pitch accent in English. *Word*, 14:109–149.
- Bolinger, D. L. (1972). Accent is predictable (if you’re a mind-reader). *Language*, 48:633–644.
- Breen, M., Fedorenko, E., Wagner, M., and Gibson, E. (2010). Acoustic correlates of information structure. *Language and Cognitive Processes*, 25(7):1044–1098.
- Bresnan, J. (1971). Sentence stress and syntactic transformations. *Language*, 47:257–280.
- Büring, D. (2007). Semantics, Intonation and Information Structure. In Ramchand, G. and Reiss, C., editors, *The Oxford Handbook of Linguistic Interfaces*, pages 445–474. Oxford University Press, New York.
- Büring, D. (2016). *Intonation and Meaning*. Oxford University Press, Oxford.
- Cambier-Langeveld, T. and Turk, A. (1999). A cross-linguistic study of accentual lengthening: Dutch vs English. *Phonetics*, 27:255–280.
- Campbell, N. and Beckman, M. (1997). Stress, prominence, and spectral tilt. *Proceedings of ESCA Workshop on Intonation: Theory, Models and Applications*, pages 67–70.
- Catford, J. C. (1977). *Fundamental Problems in Phonetics*. Edinburgh University Press, Edinburgh.
- Chafe, W. L. (1974). Language and consciousness. *Language*, 50(1):111–133.
- Chafe, W. L. (1976). Givenness, contrastiveness, definiteness, subjects, topics and point of view. In Li, C., editor, *Subject and Topic*, pages 25–56. Academic Press, New York.
- Chafe, W. L. (1994). *Discourse, Consciousness, and Time*. University of Chicago Press, Chicago.
- Chao, W. (1987). *On ellipsis*. Ph.D. thesis, University of Massachusetts, Amherst.
- Chen, M. (1970). Vowel length variation as a function of the voicing of the consonant environment. *Phonetica*, 22:129–159.
- Cho, T. and Keating, P. (2009). Effects of initial position versus prominence in English. *Phonetics*, 37:466–485.
- Cho, T. and McQueen, J. M. (2005). Prosodic influences on consonant production in Dutch: Effects of prosodic boundaries, phrasal accent and lexical stress. *Phonetics*, 33:121–157.
- Chodroff, E. and Cole, J. (2018). Information structure, affect, and prenuclear prominence in American English. *Proceedings of Interspeech*, pages 1849–1852.
- Chodroff, E. and Cole, J. (2019). The phonological and phonetic encoding of information structure in American English nuclear accents. *Proceedings of the ICPHS 2019*, pages 1570–1574.

- Chomsky, N. (1971). Deep structure, surface structure and semantic interpretation. In Steinberg, D. and Jakobovits, L. A., editors, *Semantics: An Interdisciplinary Reader*, pages 183–216. Cambridge University Press, Cambridge.
- Chomsky, N. and Halle, M. (1968). *The Sound Pattern of English*. Harper and Row, New York.
- Chomsky, N. and Lasnik, H. (1993). The Theory of Principles and Parameters. In Jacobs, J., von Stechow, A., Sternefeld, W., and Vennemann, T., editors, *Syntax: An International Handbook of Contemporary Research*, volume 1, pages 506–569. Walter de Gruyter, Berlin.
- Clark, H. H. and Haviland, S. E. (1977). Comprehension and the given-new contract. In Freedle, R. O., editor, *Discourse Production and Comprehension*, pages 1–40. Ablex Publishing Corporation, Norwood, NJ.
- Clark, H. H. and Marshall, C. R. (1981). Definite reference and mutual knowledge. In Joshi, A., Webber, B., and Sag, I., editors, *Elements of Discourse Understanding*, pages 10–63. Cambridge University Press, Cambridge.
- Cornish, F. (1999). *Anaphora, Discourse, and Understanding: Evidence from French and English*. Oxford University Press, Oxford.
- Culicover, P. and Rochemont, M. S. (1983). Stress and focus in English. *Language*, 59:123–165.
- Cutler, A. and Darwin, C. J. (1981). Phoneme-monitoring reaction time and preceding prosody: Effects of stop closure duration and of fundamental frequency. *Perception and Psychophysics*, 29:217–224.
- Dalrymple, M., Sheiber, S. M., and Pereira, F. C. N. (1991). Ellipsis and higher-order unification. *Linguistics and Philosophy*, 14(4):399–452.
- Dogil, G. and Williams, B. (1999). The phonetic manifestation of word stress. In van der Hulst, H., editor, *Word Prosodic Systems in the Languages of Europe*, pages 273–311. Mouton de Gruyter, Berlin.
- Drummond, A. (2016). Ibex 0.3.7 Manual.
- Ferreira, F., Anes, M. D., and Horine, M. D. (1996a). Exploring the use of prosody during language comprehension using the auditory moving window technique. *Journal of Psycholinguistic Research*, 25(2):273–290.
- Ferreira, F., Henderson, J. M., Anes, M. D., Weeks, Jr., P. A., and McFarlane, D. K. (1996b). Effects of lexical frequency and syntactic complexity in spoken-language comprehension: Evidence from the auditory moving-window technique. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(2):324–335.
- Fiengo, R. and May, R. (1994). *Indices and Identity*. MIT Press, Cambridge, MA.

- Flege, J. E. and Bohn, O. S. (1989). An instrumental study of vowel reduction and stress placement in Spanish-accented English. *Studies in Second Language Acquisition*, 11:35–62.
- Fletcher, J. (2010). The prosody of speech: Timing and rhythm. In Hardcastle, W., Laver, J., and Gibbon, F., editors, *The Handbook of Phonetic Sciences*, pages 523–602. Wiley-Blackwell, Oxford, 2nd edition.
- Fox, D. (1999). Focus, parallelism, and accommodation. *Proceedings of the Ninth Conference on Semantics and Linguistic Theory*, pages 70–90.
- Fox, D. (2000). *Economy and Semantic Interpretation*. The MIT Press, Cambridge, MA.
- Frank, M. C. and Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336(6084):998.
- Franke, M. (2009). *Signal to act: Game theory in pragmatics*. Ph.D. thesis, Universiteit van Amsterdam.
- Frazier, L. (2015). Two interpretive systems for natural language? *Journal of Psycholinguistic Research*, 44(1):7–25.
- Frazier, L. (2019). Ellipsis and psycholinguistics. In van Craenenbroeck, J. and Temmerman, T., editors, *The Oxford Handbook of Ellipsis*. Oxford University Press, Oxford.
- Frazier, L. and Clifton Jr., C. (2005). The syntax-discourse divide: Processing ellipsis. *Syntax*, 8(2):121–174.
- Frazier, L. and Clifton Jr., C. (2011). Dynamic interpretations: Finding an antecedent for VPE. In Harris, J. A. and Grant, M., editors, *University of Massachusetts Occasional Papers in Linguistics: Processing Linguistic Structure*, pages 21–35. GLSA, Amherst, MA.
- Fry, D. B. (1955). Duration and intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America*, 27:765–768.
- Fry, D. B. (1958). Experiments in the perception of stress. *Language and Speech*, 1:126–152.
- Fry, D. B. (1965). The dependence of stress judgments on vowel formant structure. *Proceedings of the 5th International Congress of Phonetic Sciences*, pages 306–311.
- Futrell, R. and Levy, R. (2017). Noisy-context surprisal as a human sentence processing cost model. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 688–698.
- Garellek, M. (2015). Perception of glottalization and phrase-final creak. *Journal of the Acoustical Society of America*, 137(2):822–831.
- Geiger, J. and Xiang, M. (2017). “Context can!”: Contextual accommodation in exophoric and anaphoric verb phrase ellipses. *Proceedings of the 47th Annual Meeting of the North East Linguistic Society*, 2:35–44.

- Giannakidou, A. (1998). *Polarity Sensitivity as (Non)veridical Dependency*. John Benjamins, Amsterdam.
- Gibson, E., Bergen, L., and Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, 110(20):8051–8056.
- Goodman, N. D. and Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, 5(1):173–184.
- Gramming, P., Sundberg, J., Ternström, S., Leanderson, R., and Perkins, W. (1988). Relationship between changes in voice pitch and loudness. *Voice*, 2:118–126.
- Grant, M., Clifton Jr., C., and Frazier, L. (2012). The role of non-actuality implicatures in processing elided constituents. *Journal of Memory and Language*, 66(1):326–343.
- Grice, H. P. (1975). Logic and conversation. In Cole, P. and Morgan, J., editors, *Studies in Syntax and Semantics III: Speech Acts*, pages 183–198. Academic Press, New York.
- Gundel, J. K., Hedberg, N., and Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, 69(2):274–307.
- Gunter, R. (1966). On the placement of accent in dialogue: A feature of Context Grammar. *Journal of Linguistics*, 2:159–179.
- Gussenhoven, C. (1984). Testing the reality of focus domain. *Language and Speech*, 26:61–80.
- Gussenhoven, C. (2002). Intonation and interpretation: Phonetics and phonology. *Proceedings of the First International Conference on Speech Prosody*, pages 47–57.
- Haïk, I. (1987). Bound VPs that need to be. *Linguistics and Philosophy*, 10:503–530.
- Halliday, M. (1967). Notes on transitivity and theme in English, Part 2. *Journal of Linguistics*, 3(2):199–244.
- Hankamer, J. (1978). On the nontransformational derivation of some null VP anaphors. *Linguistic Inquiry*, 9(1):66–74.
- Hankamer, J. and Sag, I. (1976). Deep and surface anaphora. *Linguistic Inquiry*, 7(3):391–428.
- Hardt, D. (1990). A corpus-based survey of VP ellipsis. Manuscript.
- Hardt, D. (1992). VP ellipsis and contextual interpretation. *Proceedings of the International Conference on Computational Linguistics (COLING-92)*, pages 303–309.
- Hardt, D. (1993). *Verb phrase ellipsis: Form, meaning, and processing*. Ph.D. Thesis, University of Pennsylvania.

- Heim, I. (1991). Artikel und Definitheit. In von Stechow, A. and Wunderlich, D., editors, *Semantik: Ein internationales Handbuch der zeitgenössischen Forschung*, pages 487–535. de Gruyter, Berlin.
- Henton, C. and Bladon, A. (1988). Creak as a sociophonetic marker. In Hyman, L. M. and Li, C. N., editors, *Language, Speech and Mind: Studies in Honour of Victoria A. Fromkin*. Routledge, London and New York.
- Hirano, M., Ohala, J., and Vennard, W. (1969). The function of laryngeal muscles in regulating fundamental frequency and intensity of phonation. *Journal of Speech, Language, and Hearing Research*, 12(3):616.
- House, A. (1961). On vowel duration in English. *Journal of the Acoustical Society of America*, 33:1174–1178.
- House, A. and Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *Journal of the Acoustical Society of America*, 25:105–113.
- Isenberg, D. and Gay, T. (1978). Acoustic correlates of perceived stress in an isolated synthetic disyllable. *Journal of the Acoustical Society of America*, 64(S1).
- Ito, K., Speer, S. R., and Beckman, M. (2004). Informational status and pitch accent distribution in spontaneous dialogues in English. *Proceedings of Speech Prosody*, pages 279–282.
- Jackendoff, R. (1972). *Semantic Interpretation in Generative Grammar*. MIT Press, Cambridge, MA.
- Jäger, G. (2011). Game-theoretical pragmatics. In van Benthem, J. F. and ter Meulen, A., editors, *Handbook of Logic and Language*, pages 467–491. Elsevier, London, 2nd edition.
- Jakobson, R. (1971). Shifters, verbal categories, and the Russian verb. In *Selected Writings II*. Mouton, The Hague.
- Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press, Oxford, 3rd edition.
- Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.
- Katzir, R. (2007). Structurally-defined alternatives. *Linguistics and Philosophy*, 30:669–690.
- Kehler, A. (1993a). A discourse copying algorithm for ellipsis and anaphora resolution. *Proceedings of the Sixth Conference of the European Chapter of the Association for Computational Linguistics*, pages 203–212.
- Kehler, A. (1993b). The effect of establishing coherence in ellipsis and anaphora resolution. *Proceedings of the 31st Conference of the Association for Computational Linguistics*, pages 62–69.

- Kehler, A. (2000). Coherence and the resolution of ellipsis. *Linguistics and Philosophy*, 23:533–575.
- Kehler, A. (2002). *Coherence in Discourse*. CSLI Publications, Stanford, CA.
- Kehler, A. (2019). Ellipsis and discourse. In van Craenenbroeck, J. and Temmerman, T., editors, *The Oxford Handbook of Ellipsis*. Oxford University Press, Oxford.
- Kehler, A. and Rohde, H. (2013). A probabilistic reconciliation of coherence-driven and centering-driven theories of pronoun interpretation. *Theoretical Linguistics*, 39(1-2):1–37.
- Kertz, L. (2008). Focus structure and acceptability in verb phrase ellipsis. *Proceedings of the 27th West Coast Conference on Formal Linguistics*, pages 283–291.
- Kertz, L. (2013). Verb phrase ellipsis: The view from information structure. *Language*, 89(3):390–428.
- Kim, C. S., Kobele, G. M., Runner, J. T., and Hale, J. T. (2011). The acceptability cline in VP ellipsis. *Syntax*, 14:318–354.
- Kitagawa, Y. (1991). Copying identity. *Natural Language and Linguistic Theory*, 9(3):497–536.
- Klatt, D. H. (1975). Vowel lengthening is syntactically determined in connected discourse. *Phonetics*, 3:129–140.
- Kochanski, G., Grabe, E., Coleman, J., and Rosner, B. (2005). Loudness predicts prominence: Fundamental frequency lends little. *Journal of the Acoustical Society of America*, 118:1038–1054.
- Kratzer, A. (1991). The representation of focus. In von Stechow, A. and Wunderlich, D., editors, *Semantics: An International Handbook of Contemporary Research*, pages 804–825. de Gruyter, Berlin.
- Krifka, M. (1998). Additive particles under stress. *Proceedings of SALT 8*, pages 111–128.
- Kuno, S. (1972). Functional sentence perspective: A case study from Japanese and English. *Linguistic Inquiry*, 3:269–320.
- Ladd, D. R. (1980). *The Structure of Intonational Meaning: Evidence from English*. Indiana University Press, Bloomington.
- Ladd, D. R. (1996). *Intonational Phonology*. Cambridge University Press, Cambridge.
- Lakoff, G. (1968). Pronouns and reference. In McCawley, J. D., editor, *Syntax and Semantics 7: Notes from the Linguistic Underground*. Academic Press, New York.
- Lakoff, G. (1972). The global nature of the nuclear stress rule. *Language*, 48:285–303.
- Lappin, S. (1992). The syntactic basis of ellipsis resolution. *Proceedings of the Stuttgart Ellipsis Workshop*.

- Lassiter, D. and Goodman, N. D. (2017). Adjectival vagueness in a Bayesian model of interpretation. *Synthese*, 194(10):3801–3836.
- Levy, R. (2008). A noisy-channel model of rational human sentence comprehension under uncertain input. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 234–243.
- Lewis, D. (1979). Scorekeeping in a language game. *Journal of Philosophical Logic*, 8:339–359.
- Lieberman, P. (1960). Some acoustic correlates of word stress in American English. *Journal of the Acoustical Society of America*, 32:451–454.
- Lobeck, A. (1999). VP ellipsis and the Minimalist Program: Some speculations and proposals. In Lappin, S. and Benmamoun, E., editors, *Fragments: Studies in Ellipsis and Gapping*, pages 98–123. Oxford University Press, Oxford.
- Luce, K. B., Geiger, J., Xiang, M., and Kennedy, C. (2018). Interpretations of VP anaphora through reference to salient events. *Proceedings of the Linguistic Society of America*, 3(38):1–13.
- Marinis, T. (2010). Using on-line processing methods in language acquisition research. In Unsworth, S. and Blom, E., editors, *Experimental Methods in Language Acquisition Research*. John Benjamins, Amsterdam.
- Maurer, G., Tanenhaus, M. K., and Carlson, G. N. (1995). A note on parallelism effects in processing deep and surface verb-phrase anaphors. *Language and Cognitive Processes*, 10:1–12.
- Mehler, J. (1963). Some effects of grammatical transformations on the recall of English sentences. *Journal of Verbal Learning and Verbal Behavior*, 2:346–351.
- Merchant, J. (2001). *The Syntax of Silence*. Oxford University Press.
- Merchant, J. (2004). Fragments and ellipsis. *Linguistics and Philosophy*, 27:661–738.
- Merchant, J. (2008). Variable island repair under ellipsis. In Johnson, K., editor, *Topics in Ellipsis*, pages 132–153. Oxford University Press, Oxford.
- Merchant, J. (2013). Voice and ellipsis. *Linguistic Inquiry*, 44:77–108.
- Merchant, J. (2018). Verb-stranding predicate ellipsis in Greek, implicit arguments, and ellipsis-internal focus. In Merchant, J., Mikkelsen, L., Rudin, D., and Sasaki, K., editors, *A Reasonable Way to Proceed: Essays in Honor of Jim McCloskey*, pages 229–269. University of California eScholarship.
- Merchant, J. (2019). Ellipsis: A survey of analytical approaches. In van Craenenbroeck, J. and Temmerman, T., editors, *The Oxford Handbook of Ellipsis*, pages 19–45. Oxford University Press, Oxford.

- Miller, P. and Hemforth, B. (2014). Verb phrase ellipsis with nominal antecedents. Manuscript.
- Miller, P. and Pullum, G. K. (2013). Exophoric VP ellipsis. In Hofmeister, P. and Norcliffe, E., editors, *The Core and the Periphery: Data-driven Perspectives on Syntax Inspired by Ivan A. Sag*. University of Chicago Press, Chicago.
- Morton, J. and Jassem, W. (1965). Acoustic correlates of stress. *Language and Speech*, 8:159–181.
- Murphy, G. L. (1985). Processes of understanding anaphora. *Journal of Memory and Language*, 24:290–303.
- Neeleman, A. and Reinhart, T. (1998). Scrambling and the PF interface. In Butt, M. and Gueder, W., editors, *The Projection of Arguments*, pages 309–353. CSLI Publications, Stanford, CA.
- Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Kastman, E., and Lindeløv, J. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, 51(1):195–203.
- Percus, O. (2006). Antipresuppositions. In Ueyama, U., editor, *Theoretical and Empirical Studies of Reference and Anaphora: Toward the Establishment of Generative Grammar as Empirical Science*. Japan Society for the Promotion of Science.
- Peterson, G. E. and Lehiste, I. (1960). Duration of syllable nuclei in English. *Journal of the Acoustical Society of America*, 32:693–703.
- Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation*. Ph.D. thesis, MIT.
- Pierrehumbert, J. B. and Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In Cohen, P., Morgan, J., and Pollack, M., editors, *Intentions in Communication*. MIT Press, Cambridge, MA.
- Prince, E. F. (1981). Toward a taxonomy of given-new information. In Cole, P., editor, *Radical Pragmatics*, pages 223–256. Academic Press, New York.
- Prince, E. F. (1992). The ZPG letter: Subjects, definiteness, and information-status. In Thompson, S. A. and Mann, W. C., editors, *Discourse Description: Diverse Analyses of a Fund Raising Text*, pages 295–325. John Benjamins, Amsterdam.
- Pullum, G. K. (2000). Hankamer was! *Jorge Hankamer WebFest*.
- Qing, C. and Franke, M. (2015). Variations on a Bayesian theme: Comparing Bayesian models of referential reasoning. In Zeevat, H. and Schmitz, H.-C., editors, *Bayesian Natural Language Semantics and Pragmatics*, pages 201–220. Springer International Publishing.

- Raphael, L. J. (1972). Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English. *Journal of the Acoustical Society of America*, 51:1296–1303.
- Rathcke, T. (2017). How truncating are ‘truncating languages’? Evidence from Russian and German. *Phonetica*, 73:194–228.
- Rietveld, A. C. M. and Gussenhoven, C. (1985). On the relation between pitch excursion size and pitch prominence. *Phonetics*, 13:299–308.
- Roberts, C. (2012). Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5(6):1–69.
- Rochemont, M. S. (1986). *Focus in Generative Grammar*. John Benjamins, Amsterdam.
- Rodrigues, C., Nevins, A., and Vicente, L. (2009). Cleaving the interactions between sluicing and P-stranding. *Romance Languages and Linguistic Theory 2006: Selected Papers from ‘Going Romance’*, pages 175–198.
- Rooth, M. (1985). *Association with focus*. Ph.D thesis, University of Massachusetts, Amherst.
- Rooth, M. (1992a). Ellipsis redundancy and reduction redundancy. *Proceedings of the Stuttgart Ellipsis Workshop*.
- Rooth, M. (1992b). A theory of focus interpretation. *Natural Language Semantics*, 1(1):75–116.
- Rosenberg, A. and Hirschberg, J. (2006). On the correlation between energy and pitch accent in read English speech. *Proceedings of Interspeech*.
- Rosner, B. S. and Pickering, J. B. (1994). *Vowel Perception and Production*. Oxford University Press, Oxford.
- Ross, J. R. (1969). Guess who? *Proceedings of the Fifth Annual Meeting of the Chicago Linguistic Society*, pages 252–286.
- Ryskin, R., Futrell, R., Kiran, S., and Gibson, E. (2018). Comprehenders model the nature of noise in the environment. *Cognition*, 181:141–150.
- Sæbø, K. J. (2004). Conversational contrast and conventional parallel: Topic implicatures and additive presuppositions. *Journal of Semantics*, 21(2):199–217.
- Sag, I. (1976). *Deletion and Logical Form*. Ph.D. thesis, MIT.
- Sag, I. and Hankamer, J. (1984). Toward a theory of anaphoric processing. *Linguistics and Philosophy*, 7:325–345.
- Sauerland, U. (2005). Don’t interpret focus! Why a presuppositional account of focus fails and how a presuppositional account of givenness works. *Proceedings of Sinn und Bedeutung*, 9:370–384.

- Schachter, P. (1977). Does she or doesn't she? *Linguistic Inquiry*, 8(4):763–767.
- Schachter, P. (1978). English propredicates. *Linguistic Analysis*, 4:187–224.
- Schmerling, S. F. (1976). *Aspects of English Sentence Stress*. University of Texas Press, Austin, TX.
- Schwarzschild, R. (1999). Givenness, AvoidF and other constraints on the placement of accent. *Natural Language Semantics*, 7(2):141–177.
- Selkirk, E. O. (1984). *Phonology and Syntax: The Relation between Sound and Structure*. MIT Press, Cambridge, MA.
- Selkirk, E. O. (1995). Sentence prosody: Intonation, stress, and phrasing. In Goldsmith, J. A., editor, *The Handbook of Phonological Theory*, pages 550–569. Cambridge University Press, Cambridge.
- Shapiro, L. P. and Hestvik, A. (1995). On-line comprehension of VP-ellipsis: Syntactic reconstruction and semantic influence. *Journal of Psycholinguistic Research*, 24:517–532.
- Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. B., and Hirschberg, J. (1992). ToBI: A standard for labelling English prosody. *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, 2:867–870.
- Silverman, K. and Pierrehumbert, J. (1990). The timing of prenuclear high accents in English. *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, pages 72–106.
- Sluijter, A. M. and van Heuven, V. J. (1996). Acoustic correlates of linguistic stress and accent in Dutch and American English. *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 630–633.
- Smith, R. and Rathcke, T. (2020). Dialectal phonology constrains the phonetics of prominence. *Phonetics*, 78.
- Snider, N. and Runner, J. T. (2014). Semantic and phonological neighbors are retrieved in verb phrase ellipsis: Evidence for syntactic structure. Manuscript.
- Stalnaker, R. (1974). Pragmatic presuppositions. In Munitz, M. K. and Unger, P. K., editors, *Semantics and Philosophy*, pages 197–214. New York University Press, New York.
- Stockwell, R. (1972). The role of intonation: Reconsiderations and other considerations. In Bolinger, D. L., editor, *Intonation*, pages 87–109. Penguin, Harmondsworth, England.
- Takahashi, S. and Fox, D. (2005). MaxElide and the re-binding problem. In Georgala, E. and Howell, J., editors, *Proceedings of SALT 15*, pages 223–240.
- Tancredi, C. (1992). *Deletion, deaccenting, and presupposition*. Ph.D. thesis, MIT.

- Tanenhaus, M. K. and Carlson, G. N. (1990). Comprehension of deep and surface verbphrase anaphors. *Language and Cognitive Processes*, 5(4):257–280.
- Terken, J. (1991). Fundamental frequency and perceived prominence of accented syllables. *Journal of the Acoustical Society of America*, 89:1768–1776.
- Thoms, G. (2015). Syntactic identity, parallelism and accommodated antecedents. *Lingua*, 166:172–198.
- Tilsen, S. (2016). A shared control parameter for F0 and intensity. *Proceedings of Speech Prosody*, pages 1066–1070.
- Townsend, D. J. and Bever, T. G. (2001). *Sentence Comprehension: The Integration of Habits and Rules*. MIT Press, Cambridge, MA.
- Turk, A. and Shattuck-Hufnagel, S. (2000). Word-boundary-related duration patterns in English. *Phonetics*, 28(4):397–440.
- Turk, A. E. and White, L. (1999). Structural influences on accentual lengthening in English. *Journal of Phonetics*, 27(2):171–206.
- Umeda, N. (1975). Vowel duration in American English. *Journal of the Acoustical Society of America*, 58:434–445.
- van Craenenbroeck, J. (2010). *The Syntax of Ellipsis: Evidence from Dutch Dialects*. Oxford University Press, New York.
- van Craenenbroeck, J. (2012). Ellipsis, identity, and accommodation. Manuscript.
- van Deemter, K. (1994). What’s new? A semantic perspective on sentence accent. *Journal of Semantics*, 11:1–31.
- van Deemter, K. (1999). Contrastive stress, contrariety, and focus. In Bosch, P. and van der Sandt, R., editors, *Focus: Linguistic, Cognitive, and Computational Perspectives*, pages 3–17. Cambridge University Press, Cambridge.
- Vicente, L. (2008). Syntactic isomorphism and non-isomorphism under ellipsis. Manuscript.
- Wagner, M. (2012). Focus and givenness: A unified approach. In Kuerov, I. and Neeleman, A., editors, *Contrasts and Positions in Information Structure*, pages 102–147. Cambridge University Press, Cambridge.
- Wasow, T. (1972). *Anaphoric relations in English*. Ph.D. thesis, MIT.
- Webber, B. L. (1978). *A formal approach to discourse anaphora*. Ph.D. thesis, Harvard University.
- Wells, J. C. (1982). *Accents of English*, volume 1. Cambridge University Press, Cambridge.

- Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., and Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America*, 92:1707–1717.
- Williams, E. (1977). Discourse and logical form. *Linguistic Inquiry*, 8:101–139.
- Xiang, M., Grove, J., and Merchant, J. (2019). Structural priming in production through ‘silence’: An investigation of verb phrase ellipsis and null complement anaphora. *Glossa*, 4(1):67.
- Xu, Y. (2013). ProsodyPro: A tool for large-scale systematic prosody analysis. *Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013)*, pages 7–10.
- Yee, E. and Sedivy, J. C. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(1):1–14.
- Yuan, J. and Liberman, M. (2008). Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America*, 123(5):3878.
- Zhang, Y. and Francis, A. L. (2010). The weighting of vowel quality in native and non-native listeners’ perception of English lexical stress. *Phonetics*, 38:260–271.