JDHCS 2010
Volume 1 Number 2

Addressing Subjectivity and Replicability in Thematic Classification of Literary Texts: Using Cluster Analysis to Derive Taxonomies of Thematic Concepts in the Thomas Hardy's Prose Fiction

A. A. Omar, School of English Literature, Language and Linguistics, Newcastle University

Thematic classification of Thomas Hardy's work has traditionally been based partly on textual content and partly on biographical considerations. These analyses and criticisms have been generated by what will henceforth be referred to as 'the philological method'; that is, by individual researchers' reading of printed materials and the intuitive abstraction of generalizations from that reading. A major problem with studies in this tradition is that they are not objective or replicable. In order to address issues of objectivity and replicability, this paper proposes an automated text clustering of the prose fiction works of Thomas Hardy using cluster analysis based on a vector space model (VSM) representation of the lexical content of the selected texts. The results reported here indicate that the proposed clustering structures yield usable results in understanding the thematic structure of Hardy's prose fiction texts and that they do so in an objective and replicable way.

The remainder of this discussion is organized as follows: part 1 is the introduction, part 2 is methodology, part 3 covers data preparation, part 4 is hierarchical cluster analysis, part 5 is an interpretation of the results, and part 6 is the conclusion.

1. Introduction

Almost all of the work on the thematic classification of Hardy's prose writings is theoretically driven. That is, classification criteria are selected by the critic based on some critical theory or framework (e.g. formal, biographical/historical, moral, Victorian, anti-Victorian, feminist, psychoanalytic, postcolonial, philosophical, religious, sociological, anthropological, etc.) supported by personal knowledge and evaluation of the texts. Moreover, many existing accounts follow the stereotypical classifications of what might be called the Hardy Critical Industry. In other words, many of Hardy's commentators are willing to agree with conventional, well-known evaluations of Hardy even though such evaluations conflict with their critical presuppositions. Two examples of this are given. First, many commentators have favored the idea of classifying Hardy's works into major and minor novels in relation to subject-matter, and many studies use that dichotomy without giving reasons for its adoption.[1] Second, many thematic reviews[2] of Hardy use the term 'Wessex novels' in reference to nine or ten of Hardy's novels without explaining why these nine or ten should texts constitute variants of the same theme apart from the fact that they are about Wessex.

The problem with such classifications is that they are neither objective nor replicable. Regardless of the adopted critical approach, thematic classifications of Hardy's work in the philological tradition are in one way or another reflections of the critics' own judgments, which can be affected by personal feelings, emotions, impressions, or prejudices. Moreover, a critic cannot set the definite criteria he used for his classification so that it can be replicated or repeated by another researcher. Even worse, it cannot guarantee that two critics following the same approach —say, the realistic or Victorian approach—will definitely reach the same conclusions. In the end it is not clear which to choose because there are no criteria except subjective ones.

---

[1] Harvey, 2003; Abercrombie, 1912.

[2] Duncan, 2002; Thomas, 1999; Lodge, 1974; Windle, 1902.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

In this context, the present discussion asks the following research question:

Can an experimentally replicable, objective, and conceptually useful classification based on empirical evidence abstracted from Thomas Hardy's prose fiction texts be defined?

Where:

- The classification is taken to be experimentally replicable if it can be duplicated by anyone using the same evidence and analytical methods. The idea is simply that if anyone repeats the classification using the same evidence and procedures, he definitely achieves the same results and reaches the same conclusions.
- The classification is taken to be objective if it is generated from empirical data by procedures that are both general in the sense that they are defined over arbitrary subject domains rather than for some specific application, and also not open to influences from any theoretical presuppositions that the researcher conducting the investigation might have.
- The classification is taken to be conceptually useful if the results enhance understanding of the subject domain.

2. Methodology

A number of computational approaches have been proposed to address the problem of objectivity and replicability in thematic generation and analysis. These are based in one way or another on document clustering theory. This is a broad framework that includes numerous methods for grouping similar texts together.[3] These include vector space clustering (VSC), latent semantic indexing (LSI), concept mining, explicit semantic analysis (ESA), and Netword. The one approach in the literature that seems theoretically most consistent with our goal, however, is VSC. This is a clustering method whereby documents are represented as vectors in a high dimensional term space with the purpose of grouping similar documents together according to their similarity or distance based on lexical content, using mathematical algorithms to compute the semantic similarity between them. This paper uses exploratory multivariate analysis (EMVA) techniques for this purpose. The idea is that this discussion is concerned with grouping texts of identical/similar themes into distinct sets, which suggests that the idea of analysis becomes a multivariate data-solving problem.[4] Moreover, using EMVA methods in VSC has proved successful in many applications.[5] EMVA encompasses numerous techniques, but for present purposes cluster analysis is the most appropriate. This is simply a multivariate mathematical technique for finding relatively homogeneous clusters of cases based on proximity measures. The rationale of using cluster analysis is that it is the most appropriate technique for organizing any collection. More importantly, cluster analysis methods are used when we do not have any prior hypotheses about the data. This serves the principle of objectivity, the ultimate concern of the research.

3. Data Preparation

In text clustering, data preparation is the key to obtaining accurate clustering performance. To achieve this, variables must be carefully selected. Data analysis should be confined to only and all the important variables that contribute meaningfully to thematic structures. That is, the data

[3] Adrian et al., 2007; Mirkin, 2005; Berry, 2004; Arabie et al., 1996; Mirkin, 1996; Hartigan, 1975.

[4] Adams, 2003.

[5] Eaton, 2007.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

matrix should be built up of only and all the important content words within the documents. The rationale of adopting content words representation is that they are strong predictors of the topic(s) or content of a document. Moreover, the experimental results of document classification indicate that content word representation gives good results in identifying the content of a document and its latent structure.[6] Equally important, most studies seem to agree that up till now content word representation has been proven to be giving much better results than any other, more sophisticated approaches to clustering.

For argument's sake, let us take the following example: assume we have a document in which the most important terms are words like *minimalism*, *phase*, *specifier*, *complement*, *head*, *terminal*, *label*, *bare*, *phrase*, *and structure*. Here, it is easy to identify its topic and content. It is about syntactic minimalism. It is, moreover, definitely different from a document whose most distinctive terms are *armed*, *political*, *conflicts*, *Parliamentarians*, *Royalists*, *civil, war*, *fight*, *support*, *king*, *first*, *second*, *supporters*, *battle*, *victory*, *Christian*, and *Protestant*, which is about the English Civil War.

On another point, the study considers content words to be indicators of semantic content. In other words, the analysis identifies all the morphological variants of a given stem as just one lexical type. It is observed that variant word forms with similar semantic conceptions can be treated as equivalent. To take an example, the words *marry*, *marries*, *married*, and *marriage* have one semantic concept which must be different from *dogs* and *cats*. The analysis thus reduces all these variant forms into just one form—presumably, *marry*.

Creating a target corpus & text pre-processing

The tradition of building a corpus for text clustering applications has always been based on the assumption that the corpus is large and representative of the research domain. Thus a relevant issue in the present context is what size the corpus should be in order to support objective and reliable generalizations about Thomas Hardy's prose fiction. The corpus on which the analysis is based consists of all the known (published and unpublished) prose fiction texts of Hardy. One requirement, however, is that the texts must be pre-processed prior to data representation. In the present case, the Hardy prose texts are reduced to lists of tokens where only content words were retained. That is, function words like determiners and prepositions were removed. 45,298 content word types were identified in this way; these are the basis for the analysis.

Data representation

Documents are represented using the vector space model (VSM). The reason for this is that it is conceptually simple as well as convenient for computing semantic similarity within documents. A data Matrix H*ij* was built in which the rows H*i* represent the documents, the columns H*j* represent the lexical type variables, and the value at the H*ij* is frequency of lexical type *j* in document *i*. The data matrix H*ij* was built out of the 45,298 variables representing the 62 texts. It is thus represented as H62, 45298. The texts were given name codes (serialized from Hardy001 to Hardy062) for identification. Each matrix row therefore represents a lexical frequency profile for the corresponding text. Because each lexical variable in the profile has a semantics, the profile gives a representation of what a text is about, what it is not about, and gradations in between. However, it was observed that the matrix H62, 45298 has some characteristics that can adversely affect the validity of clustering results. First, there are many infrequent words that are represented in the data matrix. Second, some texts are very long while others are very short. And

[6] Eaton, 2007; Soucy and Mineau, 2005; Salton et al., 1975.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

finally, the data space dimensionality is so large as to be unwieldy. These must be rectified prior to analysis.

Infrequent words

An intuitively plausible criterion for variable selection is the frequency of occurrence of textual features of interest: those that occur relatively more frequently are more likely to be more important, in some sense, than those that occur relatively infrequently.[7] The assumption is that if an author uses a word repeatedly in a document, then that document is more likely to be about what the word denotes than it is to be about the denotation of an infrequently occurring word. Based on this assumption, very infrequent words are unimportant in distinguishing documents from one another and thus can be deleted.

For an *m*-row x *n*-column matrix H in which the columns represent the variables and the rows the objects they describe, the frequency of the j^{th} column is

$$freq(H\,j) = \sum_{i=1\ldots m} H_{ij}$$

The frequencies of the columns of H62, 45298 were calculated using the above function and sorted in descending order of frequency magnitude. Variables 7,977 to 45,298 were eliminated because their frequencies are too low to be significant.

Text length normalization

The 62 texts vary substantially in length, ranging from 002 Kb to 389 Kb. These are shown in Table 1.

Title	Code	Size	Title	Code	Size
A Laodicean	hardy001	371 KB	"The First Countess of Wessex"	hardy032	042 KB
A Pair of Blue Eyes	hardy002	350 KB	"Barbara Of The House Of Grebe"	hardy033	035 KB
An Indiscretion in the Life of an Heiress	hardy003	066 KB	"The Marchioness of Stonehenge"	hardy034	014 KB
Desperate Remedies	hardy004	381 KB	"Lady Mottisfont"	hardy035	015 KB
Far from the Madding Crowd	hardy005	369 KB	"The Lady Icenway"	hardy036	011 KB
Jude the Obscure	hardy006	372 KB	"Squire Petrick's Lady"	hardy037	011 KB
Tess of the D'Urbervilles	hardy007	389 KB	"Anna, Lady Baxby"	hardy038	007 KB
The Hand of Ethelberta	hardy008	378 KB	"The Lady Penelope"	hardy039	010 KB
The Mayor of Casterbridge	hardy009	302 KB	"The Duchess of Hamptonshire"	hardy040	014 KB
The Poor Man and the Lady	hardy010	002 KB	"The Honourable Laura"	hardy041	024 KB
The Well-Beloved	hardy011	170 KB	"A Changed Man"	hardy042	018 KB
The Return of the Native	hardy012	357 KB	"The Waiting Supper"	hardy043	048 KB

[7] Rijsbergen, 1979; Luhn, 1957.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

The Trumpet-Major	hardy013	288 KB	"Alicia's Diary"	hardy044	033 KB
The Woodlanders	hardy014	369 KB	"The Grave By The Handpost"	hardy045	012 KB
Two on a Tower	hardy015	256 KB	"Enter a Dragoon"	hardy046	018 KB
Under the Greenwood Tree	hardy016	152 KB	"A Tryst At An Ancient Earthwork"	hardy047	014 KB
"The Three Strangers"	hardy017	024 KB	"What The Shepherd Saw"	hardy048	020 KB
"A Tradition of Eighteen Hundred and Four"	hardy018	007 KB	"A Committee-Man of The Terror"	hardy049	014 KB
"The Melancholy Hussar of The German Legion"	hardy019	019 KB	"Master John Horseleigh, Knight"	hardy050	013 KB
"The Withered Arm"	hardy020	028 KB	"The Duke's Reappearance"	hardy051	007 KB
"Fellow-Townsmen"	hardy021	051 KB	A Mere Interlude	hardy052	033 KB
"Interlopers At The Knap"	hardy022	030 KB	"The Romantic Adventures of a Milkmaid"	hardy053	086 KB
"The Distracted Preacher"	hardy023	053 KB	"How I Built Myself a House"	hardy054	009 KB
"An Imaginative Woman"	hardy024	025 KB	"Destiny and a Blue Cloak"	hardy055	023 KB
"The Son's Veto"	hardy025	015 KB	"The Thieves Who Couldn't Help"	hardy056	007 KB
"For Conscience' Sake"	hardy026	018 KB	"Our Exploits at West Poley"	hardy057	048 KB
"A Tragedy of Two Ambitions"	hardy027	025 KB	"Old Mrs. Chundle"	hardy058	008 KB
"On The Western Circuit"	hardy028	025 KB	"The Doctor's Legend"	hardy059	011 KB
"To Please His Wife"	hardy029	017 KB	"The Spectre of the Real	hardy060	026 KB
"The Fiddler of the Reels"	hardy030	020 KB	"Blue Jimmy: The Horse Stealer"	hardy061	008 KB
"A Few Crusted Characters"	hardy031	053 KB	"The Unconquerable"	hardy062	014 KB

Table 1. Lengths of the documents in the collection.

To assess the effect of length variation on clustering performance, a hierarchical cluster analysis of H62, 7976 is carried out.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

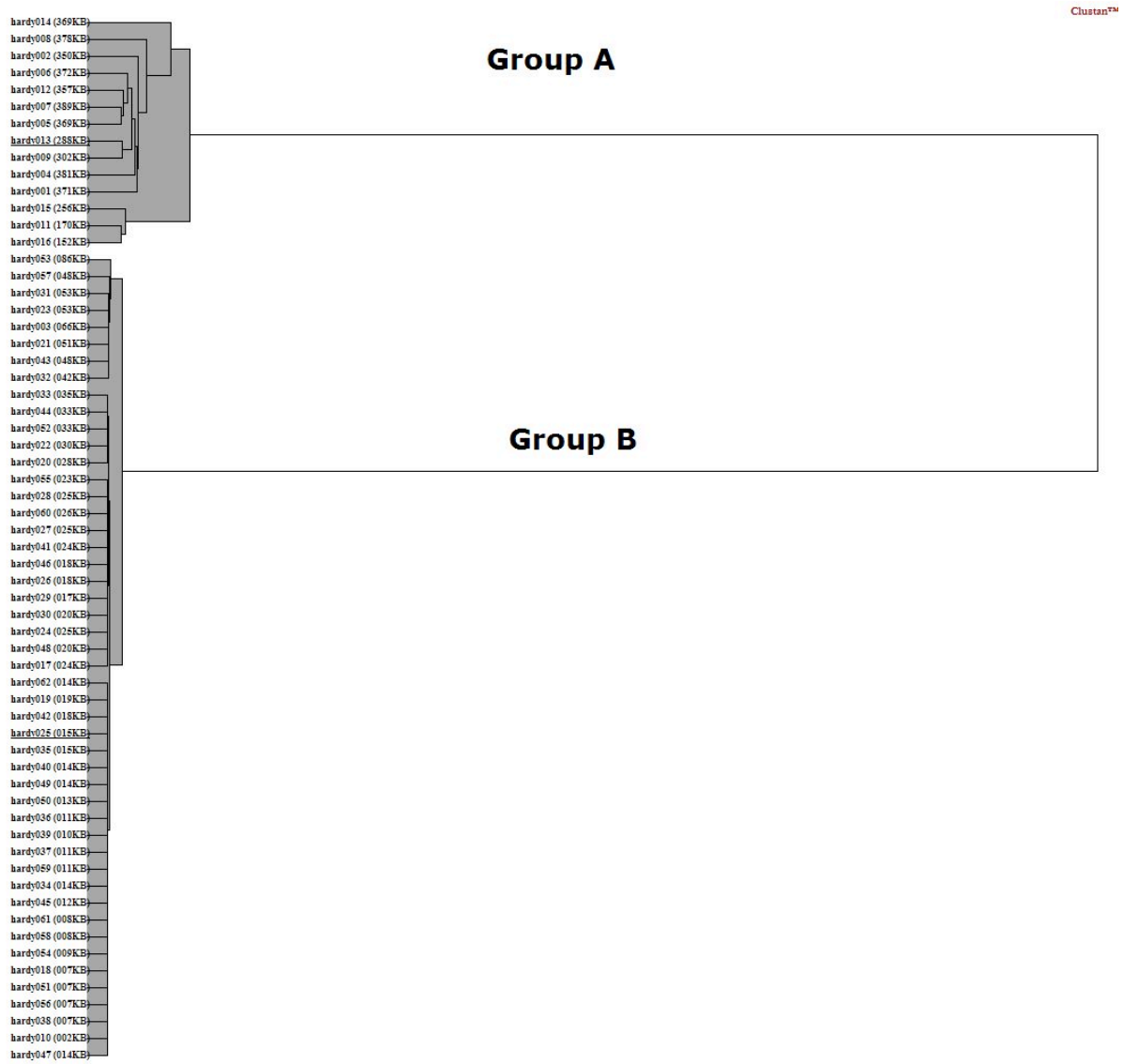

Figure 1. A hierarchical cluster analysis of H62, 7976 using squared Euclidean distance and increase in sum of squares (prior to length normalization).

The texts fall into 2 clusters: A and B. Examination of the two clusters shows that the texts do not cluster coherently in terms of thematic criteria, and the clustering in fact makes no obvious sense in terms of anything one knows about them and their subject matters. The reason for this emerges when one looks at the members on the very left of the cluster tree, each of which gives the number of bigrams in the associated text. There is a progression from the longest texts at the top of the tree to the shortest at the bottom; when correlated with cluster structure, it is easily seen that they have been clustered by length, so that A contains the longest texts and B the shortest. The idea is that in vector space, the distance between any two vectors in a space is determined by the size of the angle between the lines joining them to the origin of the space's coordinate system, and by the lengths of those lines.[8]

The problem now is that we need a clustering structure that expresses the proximities among the texts based on the semantic content, not length. To do this, the row vectors of H62, 7976 were

[8] Manning et al., 2008; Rijsbergen, 2004; Fraleigh et al., 1995; Salton, 1982; Rijsbergen, 1979.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

normalized to compensate for the variation in length among the texts so that their lexical frequency profiles could be meaningfully clustered. This normalization was relative to mean text length using the function

$$Freq(F_{ij}) = Freq(F_{ij}) \frac{\mu}{length\ (i)}$$

The effect is that the values in the vectors that represent long documents are decreased while the values of the vectors that represent short ones are increased. For documents that are near or at the mean, little or no change in the corresponding vectors took place. The overall effect is that all the corresponding documents are now in effect all the same length. A hierarchical cluster analysis of the normalized matrix is shown below, where clustering by text length is now in evidence.

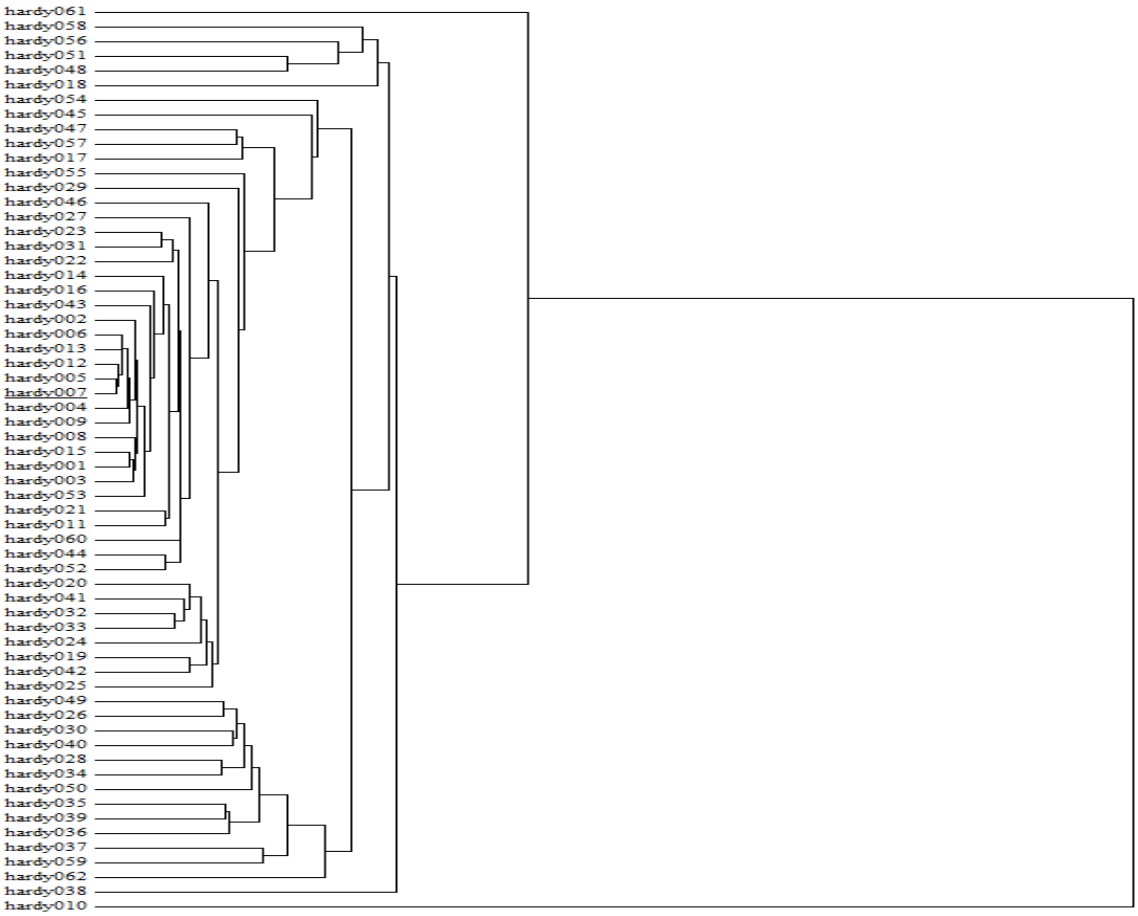

Figure 2. A hierarchical cluster analysis of the normalized H62, 7976 using Squared Euclidean distance and increase in sum of squares.

Data dimensionality

Given that data analysis must be confined to only and all the important variables, high dimensionality of text data is considered a major problem that has adverse consequences on

clustering structures. The way to overcome this problem is the removal of any superfluous variables from the dataset. To achieve this, two simple methods of dimensionality reduction were applied. These are the elimination of relatively low-variance variables and the retention of highest TF-IDF (term frequency–inverse document frequency) variables.

The elimination of low variance variables

Clustering of documents depends on there being variation in their characteristics that make them distinguished from one another. Based on this assumption, measuring the variance of the dataset becomes useful in identifying important variables.[9] The underlying principle is that high variance in a variable means that this variable varies a lot and low variance means that it does not. The assumption is therefore that high variance variables are useful in distinguishing texts, and low variance ones are not. The column vectors were sorted in descending order as shown in Figure 3: A term weighting by variance for H62, 7976.

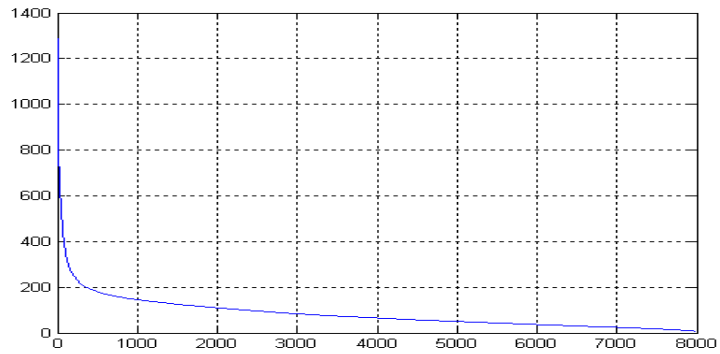

Figure 3. A term weighting by variance for H62, 7976.

Relative variance can now clearly be seen with variables of high variance on the left and variables of low variance on the right. The high-variance variables have to be kept, since they are the main criteria by which the texts can be distinguished. The flat area on the right represents the low-variance variables that contribute little or nothing to distinction among speakers, and these variables, starting about 1,001 and moving to the right, can be discarded. Variables 1001 to 7976 were eliminated because of their low variance.

The retention of the highest TF-IDF variables

TF-IDF is a statistical measure that is used to measure how important a word is to a document in a dataset matrix.[10] It has been reported to be very effective in identifying the most distinctive variables within datasets. Given that the highest TF-IDF are the most important, each column was calculated by means of TF-IDF where only the highest 400 variables were retained.

4. Hierarchical Cluster Analysis

Hierarchical cluster analysis is a two-stage procedure. The first step is the construction of a table of distances between data items by measuring the proximity among them. The choice of a measure is guided largely by the type and scale of variables and the perception of the researcher.

---

[9] Milton and Arnold, 2002; Pyle, 1999.

[10] Robertson, 2004; Spärck Jones, 1972.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

Squared Euclidean measure is used for the purpose. It uses the same equation as the Euclidean distance, but squares the standard Euclidean distance in order to place progressively greater weight on objects that are further apart. The second step is the generation of clusters based on the distance tables. There is no single best method of hierarchical clustering. Instead, the researcher selects the method that gives the most illuminative results. In the present case, this was increase in sum of squares.

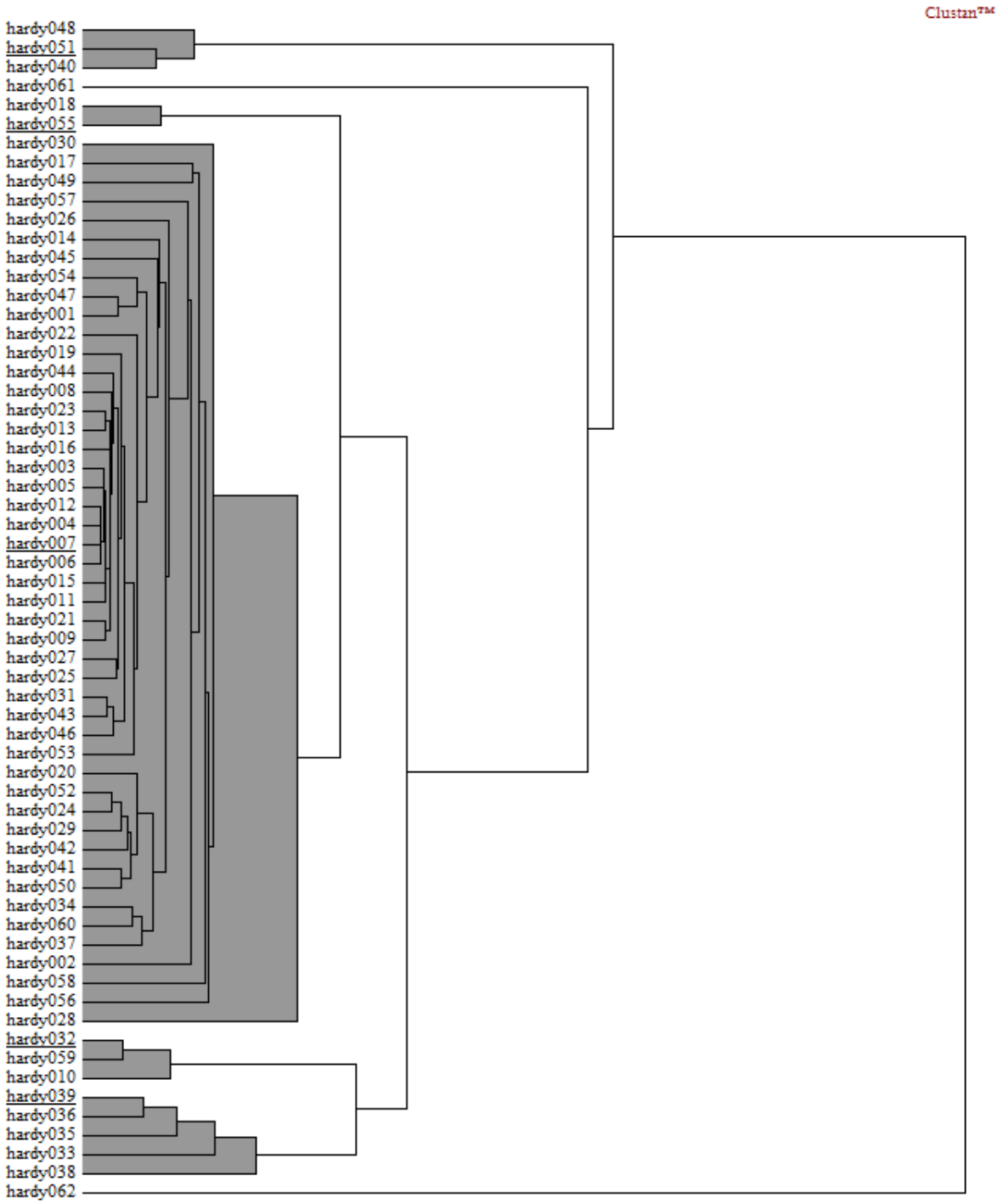

Figure 4. The hierarchical cluster analysis of H62, 400 using Squared Euclidean distance and increase in sum of squares.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

Reading the tree from the left, the document names corresponding to the row vectors are at the leaves of the tree. These are joined into clusters which are in turn combined into larger superordinate clusters, and so on recursively up the tree towards the right until the two largest clusters are amalgamated into a single cluster containing all the document row vectors. The relative lengths of the horizontal lines represent relativities of similarity between text profile groups—the longer the line, the more dissimilar the profiles. Based on this observation, it is clear that there are five main clusters, here labeled as Group 1, Group 2, Group 3, Group 4, and Group 5. These are shown in Table 2.

Cluster	Members
Group 1	hardy048 hardy051 hardy040
Group 2	hardy061
Group 3	hardy018 hardy055 hardy030 hardy017 hardy049 hardy057 hardy026 hardy014 hardy045 hardy054 hardy047 hardy001 hardy022 hardy019 hardy044 hardy008 hardy023 hardy013 hardy016 hardy003 hardy005 hardy012 hardy004 hardy007 hardy006 hardy015 hardy011 hardy021 hardy009 hardy027 hardy025 hardy031 hardy043 hardy046 hardy053 hardy020 hardy052 hardy024 hardy029 hardy042 hardy041 hardy050 hardy034 hardy060 hardy037 hardy002 hardy058 hardy056 hardy028
Group 4	hardy032 hardy059 hardy010 hardy039 hardy036 hardy035 hardy033 hardy038
Group 5	hardy062

Table 2. An illustration of the clustering structure members.

The assumption is that texts in each group have something in common that makes them similar. This does not apply to Groups 2 and 5, as each includes just one text. Group 2 includes text Hardy061, which is "Blue Jimmy: the Horse Stealer". This is one of two short stories written by Thomas Hardy in collaboration with Florence Dugdale-Hardy.[11] The story tells the adventures of Blue Jimmy, who is engaged in stealing other people's horses. This is suggested from the most distinctive columns of this cluster. These are found to be words like *horse, stealer, horse-thief*, and *case*. In the same way, Group 5 includes just one text: "The Unconquerable", the second of the collaborated short stories which Hardy wrote with Florence Dugdale-Hardy. The story is described "as a tale of mistiming and missed opportunities centering upon the love of two friends for the same woman".[12] The discussion below is thus confined to Groups 1, 3 and 4.

Interpretation

Given that the texts were clustered on the basis of lexical frequency vectors, this implies that each cluster has a characteristic lexical frequency profile which distinguishes it from the others. By comparing the lexical frequency profiles of the three clusters, therefore, it should be possible to determine the lexical items in which they differ most, and, on the basis of the lexical semantics of these items, to infer thematic characteristics of the respective clusters.

- Thematic features of Group 1

[11] Dalziel, 1992.

[12] Dalziel, 1992.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

The most distinctive vectors of this group are the words *duke, shepherd, duchess, grace, castle, dark, hut, curate, sword, battle, death, stranger, closet, struck, fear, moon, parson, beauty, thoughts, mansion,* and *midnight*. This is shown in Table 3.

Variable index	Variable name	centroid 1 frequency Group 1	centroid 2 frequency Group 3	Difference
1	duke	131923.000	411.208	131512.000
10	shepherd	21732.500	2128.120	19604.400
26	duchess	18527.600	76.513	18451.100
11	castle	17275.900	2420.390	14855.500
39	stranger	16210.000	2768.790	13441.200
25	closet	9595.210	1306.680	8288.530
46	grace	9339.780	1494.150	7845.630
32	boy	12050.400	4887.190	7163.210
87	hut	7188.690	142.033	7046.660
12	lord	10603.800	4077.520	6526.240
25	closet	9595.210	1306.680	8288.530
104	family	9486.340	3411.500	6074.840
78	wife	2042.580	8033.770	5991.180
19	captain	8372.630	2607.150	5765.480
16	curate	8582.360	3111.490	5470.870
251	sword	4016.590	114.522	3902.060
326	battle	4114.980	294.250	3820.730
168	death	6057.230	2285.580	3771.640
180	duk	3412.700	122.576	3290.120
147	mills	3234.370	22.869	3211.500
336	visitor	3577.350	483.770	3093.580
343	midnight	706.340	238.521	467.818

Table 3. The most important variables in Group 1 based on a centroid comparison between Groups 1 and 3.

The short stories included in this group are concerned with the idea of hidden or unrevealed death. This idea is repeated in the three texts, where problems of jealousy and suspicion in marriage lead to death. The main idea of all three texts is that that there is a beautiful wife who belongs to the elite. Her husband, as a man of high position, feels jealous about her and decides to take revenge against the one who is thought to be her lover because of the disgrace caused to him as a result of such an illegal relationship. This idea is tackled differently, however, in the three texts.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

- Thematic features of Group 3

The novels and short stories included in this group are most interested in the words *farmer, harnham, knight, captain, France, mop, horse, job, mademoiselle, sergeant, cloth, bit, sky, curate, rector, shore, thank, sailor, mill, river, regiment, stream, passage, trade, station, snuff, boat, mellstock, tube, cave, thieve, constable, vicar, heath, Bonaparte, landlady, college, mare, sneeze, builder, barracks, quay, tub, cabbage, village, northbrook, prisoner, cove, window, bureau, hut, two-pence, orchard, Plymouth, children, public, work, fact, sea, hill, arm, train, shop, money, arm, hope, rain, land, lane,* and *harbour*. This is shown in Table 4.

Variable index	Variable name	centroid 1 frequency Group 3	centroid 2 frequency Group 4	Difference
30	Farmer	3028.62	0	3028.62
14	Job	6720.85	391.083	2000.35
13	Horse	4773.7	8263.33	2240.2
3	Harnham	2762.9	0	6329.76
9	Knight	2624.26	0	3489.63
19	Captain	2607.15	0	2762.9
239	Stream	2350.94	435.424	1915.51
31	France	2240.2	0	2624.26
17	Mop	2000.35	0	2607.15
328	Money	1399.27	265.17	1093.52
84	Shop	1173.11	79.588	448.209
366	Cabbage	620.456	172.247	1221.92
382	Villagers	261.953	1483.87	1134.1

Table 4. The most important in Group 3 based on a centroid comparison between Groups 3 and 4.

These texts are concerned with the countryside and domestic life, social class consciousness, and romance.

- Thematic features of Group 4

The most distinctive vectors of this group are the words *squire, noble, husband, wife, marriage, love,* and *heiress*. These are shown in Table 5.

Variable index	Variable name	centroid 1 frequency Group 3	centroid 2 frequency Group 4	Difference
2	squire	1565.440	36412.900	34847.400
27	Husband	10630.900	45350.900	34720.000
78	wife	8033.770	17655.000	9621.200

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

127	noble	501.900	4360.500	3858.600
65	Marry	4348.200	7611.390	3263.190
104	family	3411.500	8806.550	5395.050
51	love	8995.130	13559.300	4564.220
77	heiress	30.036	988.326	958.290

Table 5. The most important variables in Group 1 based on a centroid comparison between Groups 3 and 4.

The texts included here are concerned with ideas of filial antagonism and disobedience, incompatible union or marriage, elopement, family disgrace, and feminine submissiveness, and humiliation.

6. Conclusion

Returning to the question posed at the beginning of this study, it is now possible to state that the novels and short stories can be thematically clustered using objective and replicable methods. Although the results agree to some degree with the accepted tradition and practice of traditional literary studies, this is done through objective methods with clear criteria, rejecting at the same time the idea of critical stereotype of thematic criticism of Hardy's prose fiction.

References

Abercrombie, L. 1912. *Thomas Hardy. A critical study.* London: Martin Secker.

Adams, R. 2003. Perceptions of innovations: exploring and developing innovation classification. PhD diss., Cranfield University.

Adrian, K., D. Stphane, and G. Tudor. 2007. Semantic clustering: Identifying topics in source code. *Information and Software Technology* 49 (3): 230-243.

Arabie, P., L. J. Hubert, and G. d. Soete. 1996. *Clustering and classification.* London: World Scientific.

Berry, M. W., ed. 2004. *Survey of text mining: Clustering, classification, and retrieval.* New York: Springer.

Dalziel, P., ed. 1992. *Thomas Hardy: The excluded and collaborative stories.* Oxford: Clarendon Press.

Duncan, I. 2002. The provincial or regional novel. In *A companion to the Victorian novel,* ed. P. Brantlinger and W. Thesing. Oxford: Blackwell.

Eaton, M. L. 2007. *Multivariate statistics: A vector space approach.* Beachwood, OH: IMS.

Fraleigh, J. B., R. A. Beauregard, and V. J. Katz. 1995. *Linear algebra.* Reading, MA: Addison-Wesley.

Hartigan, J. A. 1975. *Clustering algorithms.* New York: Wiley.

Harvey, G. 2003. *The complete critical guide to Thomas Hardy*. London: Routledge.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

Lodge, D. 1974. 'Thomas Hardy and cinematographic form. *NOVEL: A Forum on Fiction* 7 (3): 246-254.

Luhn, H. 1957. A statistical approach to mechanised encoding and searching of library information. *IBM Journal of Research and Development* 1: 309-317.

Manning, C. D., P. Raghavan, and H. Schütze. 2008. *An introduction to information retrieval.* Cambridge: Cambridge University Press.

Milton, J. S. and J.C. Arnold. 2002. *Introduction to probability and statistics: Principles and applications for engineering and the computing sciences*. New York: McGraw-Hill.

Mirkin, B. 2005. *Clustering for data mining: A data recovery approach*. Taylor & Francis Group, LLC.

Mirkin, B. G. 1996. *Mathematical classification and clustering*. London: Kluwer Academic.

Pyle, D. 1999. *Data preparation for data mining*. San Francisco, CA: Taylor & Francis.

Rijsbergen, C. J. V. 1979. *Information retrieval.* London: Butterworth.

Rijsbergen, C. J. V. 2004. *The geometry of information retrieval.* Cambridge: Cambridge University Press.

Robertson, S. 2004. Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation* 60 (5): 503-520.

Salton, G. 1982. *Introduction to modern information retrieval.* New York: McGraw-Hill.

Salton, G., A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM* 18 (11): 613–620.

Soucy, P. and G. W. Mineau. 2005. Beyond TFIDF weighting for text categorization in the vector space model. *Proceedings of the 19th International Joint Conference on Artificial Intelligence IJCAI:* 1130-1135.

Spärck Jones, K. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28: 11-21.

Thomas, J. 1999. *Thomas Hardy, femininity and dissent: Reassessing the minor novels*. New York: Macmillan.

Windle, B. C. A. 1902. *The Wessex of Thomas Hardy.* London: J. Lane.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago