Semantic Variation in Idiolect and Sociolect: Corpus Linguistic Evidence from Literary Texts

Running head: Semantic Variation in Idiolect and Sociolect

Article type: full length article

Max Louwerse

Department of Psychology / Institute for Intelligent Systems

The University of Memphis

202 Psychology Building

Memphis, TN 38152

Phone: (901) 678-2143

Fax: (901) 678-2579

Email: mlouwers@memphis.edu

Semantic Variation in Idiolect and Sociolect: Corpus Linguistic Evidence from Literary Texts

MAX M. LOUWERSE
Department of Psychology / Institute for Intelligent Systems (mlouwers@memphis.edu)

Abstract. Idiolects are person-dependent similarities in language use. They imply that texts by one author show more similarities in language use than texts between authors. Sociolects, on the other hand, are group-dependent similarities in language use. They imply that texts by a group of authors, for instance in terms of gender or time period, share more similarities within a group than between groups. Although idiolects and sociolects are commonly used terms in the humanities, they have not been investigated a great deal from corpus and computational linguistic points of view. To test several idiolect and sociolect hypotheses a factorial combination was used of time period (Modernism, Realism), gender of author (male, female) and author (Eliot, Dickens, Woolf, Joyce) totaling 16 corresponding literary texts. In a series of corpus linguistic studies using Boolean and vector models, no conclusive evidence was found for the selected idiolect and sociolect hypotheses. In final analyses testing the semantics within each literary text, this lack of evidence was explained by the low homogeneity within a literary text.

Keywords. author identification, coherence, computational linguistics, content analysis, corpus linguistics, idiolect, latent semantic analysis, literary period, sociolect.

1. Introduction

Writers implicitly leave their signature in the document they write, groups of writers do the same. *Idiolects* are similarities in the language use of an individual, *sociolects* similarities in the language use of a community of individuals. Although various theoretical studies have discussed the notion of idiolects and sociolects (Eco, 1977; Fokkema and Ibsch, 1988; Jakobson, 1987; Lotman, 1977) and those theories are widely accepted in fields like literary criticism (Fokkema and Ibsch, 1988), semiotics (Eco, 1977; Sebeok, 1991) and sociolinguistics (Wardhaugh, 1998), hypotheses derived from those theories have not often been empirically tested. The present study will test some of these hypotheses, using different computational corpus linguistic methods.

2. Idiolects, sociolects and literary periods

Both idiolect and sociolect depend on the linguistic code the writer uses. On top of this linguistic code other codes (e.g. narrative structures) can be built (Eco, 1977; Jacobson, 1987; Lotman, 1977). These complementary linguistic codes allow for texts to be culturalized. The best examples of these culturalized texts are artistic texts. These texts are thus secondary modeling systems made accessible by the primary (linguistics) modeling system. What is so special about aesthetic texts is that the author will try to deviate from currently accepted codes. By deviating from the norm texts become aesthetic. This way the deviation gradually becomes the norm of a group and by deviating from the established norm new aesthetic texts will deviate (Martindale, 1990).

In practice it is very difficult to determine these multiple encodings. On the one hand, to determine the idiolect or sociolect from a literary text, one has to look at the complementary language codes. On the other, however, the product of the multiple modeling systems is just one linguistic system. Fokkema and Ibsch (1988) argue that although the text usually doesn't yield data about complementary language codes, we are likely to find differences in the language code by comparing texts with different complementary codes (e.g. the time period). In other words, on the one hand a top-down approach could analyze those texts that share certain aspects (e.g. time of first publication) and report their similarities. On the other hand, a bottom-up approach could compare linguistic codes of different texts, and report predictions about the idiolects and sociolects. The current study will use both.

We start with the top-down approach, following Fokkema and Ibsch's (1988) theory of Modernist conjectures. According to Fokkema and Ibsch historical developments change the way we think and hence will likely have an impact on the cultural system. For instance, historical events around WWI led to principal political changes and psychological and scientific depression. Similarly, WWII created another break in world history and in our thinking. It is therefore not surprising that Fokkema and Ibsch distinguish two literary periods on the basis of these historical breaks. The first ranges from approximately 1850 to 1910 and is called Realism. The second ranges from approximately 1910 to 1940 and is called Modernism (see also Wellek and Warren, 1963).

By analyzing a number of literary texts written during this 30-year time frame, Fokkema and Ibsch are able to define a Modernist code. This code is a selection of the syntactic, pragmatic, and semantic components of the linguistic and literary options the author has available. The semantic component receives by far most attention in their study. The Modernist semantic code consists of three central semantic fields: *awareness*, *detachment* and *observation*. These fields can be visualized as concentric circles that form a first semantic zone. The field *awareness* consists of words like *awareness* and *consciousness*. The semantic field of *observation* consists of words like *observation*, *perception* and *window*. Finally, *detachment* consists of words like *depersonalization* and *departure*. In addition to this first zone of semantic fields a second zone can be distinguished. This zone contains neutral semantic fields related to the idiolect of the author. A third zone, finally, contains semantic fields that are at the bottom of the Modernist semantic hierarchy, including *economy*, *industry*, *nature*, *religion*, *agriculture*. In addition, fields like *criminality*, *psychology*, *science*, *sexuality* and *technology* that were already present in pre-Modernist literature are expanded in Modernist texts.

Throughout their study Fokkema and Ibsch show that texts written by authors who wrote literary texts in the period 1910-1940 share the pragmatic, syntactic and semantic components of the Modernist code. The notion of Modernist code has various implications. First of all, it assumes that those texts written within the Modernist time frame (e.g. 1910-1940) share particular language features, including a prominent role for the selected semantic fields. Secondly, the notion of Modernist code implies that those literary texts written within a certain time frame share particular language features (period code). Thirdly, groups of authors share language features (what we earlier called *sociolect*) that could be defined in different ways: chronologically as Fokkema and Ibsch did, but other ways are also possible. For instance, we could group authors by gender. Finally, if groups of authors share language features, texts written by an individual author must share language features (what we earlier called *idiolect*).

Accordingly we can formulate four hypotheses: 1) an *idiolect hypothesis* that predicts that linguistic features in texts by one author should not significantly differ from each other, whereas those from texts by different authors should; 2) *a sociolect-gender hypothesis*[1] that predicts that linguistic features of texts written by male authors should not significantly differ, but they should differ from texts written by female authors; 3) *a sociolect-time hypothesis* predicting that texts written within a particular time frame should not differ, but texts between time-frames should; 4) *a Modernist-code hypothesis* that predicts that Modernist texts should not only show homogeneity and differ from Realist texts, but they should also show a higher frequency of certain semantic fields. It needs to be kept in mind though that these hypotheses are stated according to a stringent criterion. For instance, it is of course possible for one author to shift in style between different periods (Watson, 1994). In the first experiment, the four hypotheses are tested using the frequency of semantic fields occurring in a series of literary texts.

3. Study 1: Semantic field comparisons using a Boolean model

Fokkema and Ibsch (1988) suggest a word frequency analysis to test the Modernist-code hypothesis. In our first study this generally accepted corpus linguistic method is used, by taking word frequency as a measure of semantic distinction. Such method can be identified as a Boolean model (Baeza-Yates and Ribeiro-Neto, 1999). This model has very precise semantics using a binary decision criterion. It is the most commonly used method in content analysis and has been extensively used in corpus linguistics in general (Biber, 1988), in social psychology (Pennebaker, 2002) and in literary studies in particular (see Louwerse and Van Peer, 2002). The four hypotheses outlined in the previous section (Modernist-code hypothesis, the sociolect-gender hypothesis, sociolect-time hypothesis and the idiolect-hypothesis) will be tested using the frequency of words in each of the semantic fields identified by Fokkema and Ibsch (1988).

3.1 MATERIALS

A total of sixteen texts were selected for the analysis following a 2 (literary period) x 2 (gender) x 4 (texts per author) design. The selection of authors followed Fokkema and Ibsch (1988). At the same time the choice of authors and texts was constrained by the availability of electronic versions of these texts (hence the focus on English texts only) and the preferred design (four corresponding texts from one author in each cell). Fokkema and Ibsch (1988, p. 192, p. 203) consider George Eliot and Charles Dickens as representatives for Realist authors. For the literary period Modernism Virginia Woolf and James Joyce were selected (Fokkema and Ibsch, 1988, p. 10). Table 1 gives an overview of the sixteen texts classified by period and gender, indicating year of publication and number of words. Despite the various text archive initiatives (e.g. Project Gutenberg, The Oxford Text Archive, The Online Books Page) finding electronic versions of texts from authors discussed in Fokkema and Ibsch (1988) and finding four texts from each author remains a daunting task. Rather than being seen as the final complete set of corpora, the sixteen selected texts should be considered as a representative sample to study the relevant research questions.

3.2 SEMANTIC FIELDS

All thirteen semantic fields Fokkema and Ibsch identify as characteristic for Modernist texts were used in this study: *consciousness*, *observation*, *detachment*, *agriculture*, *criminality*, *economy*, *industry*, *nature*, *psychology*, *religion*, *science*, *sexuality* and *technology*.

Two graduate students in cognitive psychology populated the thirteen semantic fields with lemmata. A total of 592 lemmata were created from two sources. Roget’s thesaurus was the source for the majority of the lemmata (59%). By selecting each of the semantic fields as a keyword in the thesaurus, large numbers of semantically related words were found. A second source was the WordNet database (41% of the lemmata), a large semantic network of nouns and verbs (Fellbaum, 1998). By using the label of the semantic field as a hypernym in WordNet, all related hyponyms were selected.

Obviously, in a Boolean model where precise semantics is crucial the actual word form is essential and lemmata alone do not suffice. Therefore, for each of the 592 lemmata corresponding derivations and inflections were generated, resulting into a total of 1461 word forms.

Table 1 Overview of 16 corpora used

Period	**Gender**	**Author**	**Texts**	**Year of publication**	**Number of words**
Realism	Female	George Eliot	Silas Marner	1861	75,632
			Brother Jacob	1860	20,863
			Middlemarch	1872	322,594
			Mill on the Floss	1860	214,441
	Male	Charles Dickens	Oliver Twist	1838	162,025
			Tale of Two Cities	1859	140,389
			David Copperfield	1850	363,323
			Pickwick Papers	1836	304,907
Modernism	Female	Virginia Woolf	Mrs. Dalloway	1925	81,550
			The Waves	1931	80,236
			Orlando	1928	83,562
			To the Lighthouse	1927	73,300
	Male	James	Exiles	1918	31,067

		Joyce	Dubliners	1914	71,790
			Portrait[2]	1916	90,086
			Ulysses	1922	271,722

3.3 RESULTS AND DISCUSSION

To account for different text sizes, a normalization procedure transformed the raw frequency to a basis per 1000 words of a text (Biber, 1988). The four hypotheses (idiolect, sociolect-gender, sociolect-time, and Modernist-code) were then tested on the frequency of the semantic fields in each of the sixteen texts. First, it needs to be established whether there are differences between all sixteen texts. If there are not, aggregating across authors, gender or time period would be futile. As predicted differences were found between the texts (H = 234.951, df = 15, p < .001, N = 23376). To test the idiolect-hypothesis, between-groups comparisons (texts between authors) as well as within-groups comparisons (texts by one author) were made. As predicted by the idiolect-hypothesis, the frequency of semantic fields indeed differed between authors (H = 18.49, df = 3, p < .001, N = 23376). A Mann Whitney U pair wise analysis, however, showed that this difference was due to comparing Dickens with Eliot, Woolf or Joyce (U > 0.01, Z = -3.361, p < .001, N = 11688), whereas the idiolect hypothesis predicts that differences would occur between all authors. No significant differences were found between Eliot and Joyce, Eliot and Woolf or Woolf and Joyce. Identical results were obtained for those semantic fields limited to the first zone (*awareness*, *observation*, *detachment*). The predicted between-groups differences should be accompanied by a lack of within-group differences. However, within-groups comparison showed that texts by Eliot, Woolf and Joyce differed in frequency of semantic fields (all Hs > 34.47, df = 5844, p < .001). Only the texts by Dickens confirmed the idiolect hypothesis with no differences in the frequency of the semantic fields, resulting in only very limited support for the idiolect-hypothesis.

Contrary to what was predicted by the sociolect-gender hypothesis, no differences were found between the female authors (Eliot, Woolf) and the male authors (Dickens, Joyce). Moreover, significant effects were found both within the male authors and female authors (*H* > 107.389, *df* = 7, *p* < .01, *N* = 11688), suggesting a lack of homogeneity in gender and falsifying the sociolect-gender hypothesis. When the analysis only took into account the first zone semantic fields, an effect between the gender of authors was found (*U* = .01, *Z* = -2.550, *p* = .011, *N* = 9184). Although this would support the sociolect-gender hypothesis, no homogeneity within male authors and female authors was found (*H* > 32.118, *df* = 7, *p* < .01, *N* = 4592).

For sociolect-time hypothesis a difference was found between the Realist texts and the Modernist texts (*U* = .01, *Z* = -4.076, *p* < .001, *N* = 23376). Nevertheless, as with the sociolect-gender hypothesis, this support for the sociolect-hypothesis would only be meaningful if there is homogeneity in the frequency of the semantic fields between the texts within a period. But in both the Realist texts and the Modernist texts, differences between texts were found (*H* > 79.351, *df* = 7, *p* < .001, *N* = 11688).

The lack of homogeneity in Realist texts on the one hand and in Modernist texts on the other, falsifies the sociolect-time hypothesis. Consequently, no conclusive support is found for the Modernist-code hypothesis, despite the fact that the significant difference between Realist texts and Modernist texts does show the predicted pattern. A higher frequency of the semantic fields is found in Modernist texts (*Mean* = .0023, *SE* = .001) than in Realist texts (*Mean* = .00247, *SE* = .001) but this pattern is supported by the Dickens and Woolf texts only and not by the other texts. In fact, the frequency of semantic fields in Eliot is almost as high as the frequency of fields in Woolf. Similarly, frequency of fields in Dickens is almost as high as the frequency of fields in Joyce.

Although Fokkema and Ibsch's hypotheses have strictly been followed, one could argue that the choice of the texts and contents of the semantic fields distorts the picture. To account for this possibility each of the sixteen texts was split into two halves and each of the halves were compared using a Wilcoxon Signed Ranks test. No significant difference was found for any of the texts, except for Eliot's *Middlemarch* (z = -4.47, p < .001; N = 1461) Dickens' *Copperfield* (z = -7.014, p < .001; N = 1461) and Joyce's *Ulysses* (z = -6.853, p < .001; N = 1461). There is the option of removing these texts from the analysis. However, given the importance of these texts for their respective categories, the importance of equal cell sizes and the difficulties in finding electronic versions of the required texts, we have to run the risk of making a Type II error in this study.

What can be concluded so far? Should all four hypotheses be abandoned because of a lack of evidence from the semantic field frequencies between the corpora? One problem in this study is the method. One of the obvious drawbacks of a Boolean model is its precise semantics (see Baeza-Yates and Ribeiro-Neto, 1999). The binary decision criterion means that if a word form is not found in the exact format as specified it will return a null result. It is feasible however that the semantic field is generally present in a paragraph rather than in the form of an exact string-match. The paragraph would then semantically approach the semantic field without a specific word matching the keyword. Similarly, a field might be present in the text but only by numbers of words loosely associated with the keywords used for the population of the semantic field. In other words, some kind of semantic grading scale is desirable. This is what is investigated in the second study.

4. Study 2: Semantic field comparisons using a vector model

To overcome the limitations of binary decision making (Boolean model), degrees of similarities between the selected semantic fields and texts were measured using a vector model. One of the vector models commonly used in computational linguistics is latent semantic indexing (LSI), also called latent semantic analysis (LSA).

LSA is a statistical, corpus based, technique for representing world knowledge. It takes quantitative information about co-occurrences of words in paragraphs and sentences and translates this into an *N*-dimensional space. Generally, the term ‘document’ is used for these LSA units (paragraphs or sentences), but to confuse terminology, we will use ‘text units’ here. Thus, the input of LSA is a large co-occurrence matrix that specifies the frequency of each word in a text unit. LSA maps each text unit and word into a lower dimensional space by using singular value decomposition. This way, the initially extremely large co-occurrence matrix is typically reduced to about 300 dimensions. Each word now becomes a weighted vector on *K* dimensions. The semantic relationship between words can be estimated by taking the dot product (cosine) between two vectors. What is so special about LSA is that the semantic relatedness is not (only) determined by the relation between words, but also by the words that accompany a word (see Landauer and Dumais, 1997). In other words, terms like *consciousness* and *mind* will have a high cosine value (are semantically highly related) not because they occur in the same text units together, but because words that co-occur with one equally often co-occur with the other (see Baeza-Yates and Ribeiro-Neto, 1999, Landauer and Dumais, 1997; Landauer, Foltz and Laham, 1998).

The method of statistically representing knowledge has proven to be useful in a range of studies. It has been used as an automated essay grader, comparing student essays with ideal essays (Landauer, et al., 1998). Similarly, it has been used in intelligent tutoring systems, comparing student answers with ideal answers in tutorials (Graesser, et al., 2000). LSA can measure the coherence between successive sentences (Foltz, et al., 1998). It performs as well as students on TOEFL (test of English as a foreign language) tests (Landauer and Dumais, 1997) and can even be used for understanding metaphors (Kintsch, 2000). In this second study we therefore used the populated semantic fields and compared them not to the texts as in study 1, but to the semantic LSA spaces of those texts.

4.1 MATERIALS

The same sixteen texts from the authors Eliot, Dickens, Woolf and Joyce were used. For each text a semantic space was created using the default of 300 dimensions (see Graesser, et al., 1999; for the most recent view see Hu, et al., 2003). The weighting for the index terms was kept to the default log entropy. Similarly, the default feature of disregarding common words like functional items was used. The size of the text units was generally kept at paragraphs, except in the case for dialogs when lines were chosen as text unit size, with the size of each semantic space ranging from 600 text units to 1700 text units per text.

4.2 SEMANTIC FIELDS

The same thirteen semantic fields were used as in the first study with the same population of lemmata (N = 592) and word forms (N = 1461).

4.3 RESULTS AND DISCUSSION

After the LSA spaces per text were created, the 1461 words forms for the thirteen semantic fields were compared with the LSA space, resulting in a cosine value between 0 and 1. The very large number of data points (i.e., number of word forms x the number of text units within each text) called for a more manageable analysis.

Therefore, a sample of 65,000 data points per LSA output file were randomly selected using a simple random sampling technique.

The idiolect hypothesis predicted significant differences between texts from different authors, but no significant differences between the texts of one author. As predicted, between-author groups differed from each other (*F*(1, 1040000) = 31.82, *p* < .001), with cosine values highest for Eliot (*Mean Cosine* = .040, *SD* = .059) and Dickens (*Mean Cosine* = .040, *SD* = .044), lowest for Woolf (*Mean Cosine* = .012, *SD* = .056), with Joyce in between (*Mean Cosine* = .039, *SD* = .06). However, contrary to this prediction, texts written by Eliot showed significant differences between them (*F*(3, 260000) = 4.72, *p* < .003), as did texts by Dickens (*F*(3, 260000) = 10.16, *p* < .01) and Joyce (*F*(1, 260000) = 11.49, *p* < .001). Only the texts by Woolf seemed to be more homogeneous (*F*(1, 260000) = 2.61, *p* = .05).

The sociolect-gender hypothesis predicted that texts by male authors would differ from those by female authors, whereas no differences were predicted between texts within each of these two groups. Indeed, a difference was found between these two author groups (*F* = 1, 1040000) = 5.15, *p* = .023), with higher cosine values for female authors (*Mean Cosine* = .039, *SD* = .050) than for male authors (*Mean Cosine* = .029, *SD* = .059). However, between the texts within each of the groups significant differences were also found (Male: *F*(1, 520000) = 62.28, *p* < .001), Female: *F*(1, 520000) = 31.23, *p* < .001).

The sociolect-period hypothesis predicted that no differences would be found between the texts within a period. Cosine values between texts of the Realist authors indeed did not show a difference (*p* = .5), but contrary to what was expected values between Modernist texts did show significant differences (*F*(1, 520000) = 8.67, *p* = .003). In addition, as predicted, differences between the two time periods were found (*F*(1,1040000) = 89.16, *p* < .001). However, whereas the Modernist-code hypothesis predicted that the values for the semantic fields would be higher in the Modernist texts than in the Realist texts, an opposite effect is found with higher cosine values for the Realist texts (*Mean Cosine* = .040, *SD* = .051) than the Modernist texts (*Mean Cosine* = .026, *SD* = .059). In fact, this effect can be found for all possible interactions between the Realists and Modernist texts. Similar to the findings in the previous study identical results were found for the core set of three semantic fields (*consciousness*, *observation*, *detachment*) as for the overall set of semantic fields.

In this second study similar results were found as study 1. Comparing the semantic fields to the LSA spaces of the texts rather than to the texts themselves allowed for a degree of similarity, but again results showed both between as well as within group differences. The only exception to this was the Realist texts not showing differences within the group itself. Although it is difficult to draw conclusions about the Modernist-code hypothesis without confirmation of the idiolect and sociolect hypotheses, the Modernist-code effect that was found to show an effect that did not match the prediction, with a higher average cosine value for Realist texts than for Modernist texts.

In any case, no unambiguous evidence was found for any of the four hypotheses. This would suggest a lack of empirical support for the claims made by Fokkema and Ibsch (1988). However, it is still possible that the idiolect and sociolect hypotheses hold and that only the Modernist-code hypothesis should be rejected, because of the selected semantic fields. In other words, we might still be able to find semantic similarities between groups of texts (idiolect, sociolect) but these similarities might not be contingent on the semantic fields. The idiolect and sociolect hypotheses may then be falsified by a particular selection of semantic fields, but not by the full semantic space of the texts. This option is what is explored in a third study.

5. Study 3: Between-text comparisons using a vector model

Instead of comparing a predefined list of semantic fields to words in each of the corpora (study 1) or the semantic spaces of those corpora (study 2), LSA spaces of each text were compared with each other. In other words, each text unit (paragraph or sentence) in each text was compared with each text unit (paragraph or sentence) of another text, resulting in a cosine value for each comparison. The higher the cosine value, the more similar the text units are (ranging from 0 to 1). According to the idiolect hypothesis the semantic universes of texts by one author do not show differences, whereas the semantic universes of texts between authors do. Similarly, within-gender or within-time texts are expected not to differ, but between-gender or between-time texts are. Similarities in cosine values between texts indicate homogeneity of the content. In addition, high cosine values are indicators of semantic similarities.

5.1 MATERIALS

The same LSA spaces of the sixteen texts were used as those created for the second study.

5.2 RESULTS AND DISCUSSION

In this study (LSA spaces of) texts were compared to other texts instead of to a word list as in the first studies, resulting in 256 (16 x 16) sets of cosines representing the semantic relationship between texts. A comparison of the author-matching texts with the author-non-matching texts showed differences in all four cases (All *F*s (1, 6000000) ≥ 251250, *p* < .001). When the cosine values (indicating similarity in content) were compared per idiolect, texts by Dickens differed more between themselves than between texts from other authors. The same is true for texts by Joyce. In other words, only for half of the authors (Eliot and Woolf) the author-matching texts had higher average cosine values than the author-non-matching texts.

In order to test the sociolect-gender hypothesis, texts by male authors were compared with the texts by female authors. Differences between groups were found, suggesting evidence for the sociolect-gender hypothesis (*F*(1, 3640000) = 392989.6 < .001). However, as with the unpredicted results in the idiolect hypothesis, significant differences were also found within each gender group (All *F*s (1, 1820000) ≥ 31747.8, *p* < .001). Overall, texts by female authors had a higher average cosine value, suggesting a resemblance in content, than texts by male authors (female: *Mean Cosine* = .135, *SD* = .121; male: *Mean Cosine* =.058, *SD* = .103).

As predicted by the sociolect-period hypothesis, significant differences were found between Realist-matching texts versus Modernist-matching texts (*F*(1, 3640000) = 12246.763, *p* < .001). But again, unexpected differences were also found within each period (All *F*s (1, 1820000) ≥ 1579.459, *p* < .001). Interestingly, average cosine values were higher for Realist text than for Modernist texts (Realist: *Mean Cosine* = .107, *SD* = .128; Modernist = .093, *SD* = .110). This suggests that despite the fact that there are differences between the Realist texts, they are semantically more similar to each other than Modernist texts are.

In sum, for some authors (Eliot and Woolf) similarity in content can be found, supporting an idiolect hypothesis. For other authors (Dickens and Joyce) texts differ within one author. This finding is even more interesting when we look at the sociolect-gender hypothesis. Texts by female authors show more similarities than texts by male authors. Similarities in content were also found in the sociolect-time hypothesis: In both Realist and Modernist texts more similarities were found between the texts within a period than between periods. Furthermore, Modernist texts show a greater diversity when compared to each other than Realist texts, suggested by the lower cosine values for the former compared to the latter.

6. Study 4: Within-text comparisons using a vector model

Up to now, we have found no conclusive evidence for the idiolect-hypothesis or either of the sociolect-hypotheses. Should we therefore abandon all four hypotheses? So far we have assumed that there is homogeneity in the semantics within a text. This has largely been supported by the within-text analysis when the two text halves of each texts were compared. The question however is two what extent this assumption is correct. It might be the case that within a text there are differences between the semantics. If that is the case, the lack of evidence for the idiolect and sociolect hypotheses might be explained by the heterogeneity within a text. At the same time, the hypotheses can be tested by comparing homogeneity values by author, gender and period. For this purpose an LSA analysis was carried comparing each text unit (i.e. paragraph or sentence) to every other text unit (paragraph or sentence) within a text. If texts are generally semantically consistent (the content of the text units in the text is similar), higher cosine values will be found. Texts that differ in the semantics, and are therefore semantically inconsistent, will have lower cosine values.

6.1 MATERIALS

The same LSA spaces of the literary texts from the second and third studies were used.

6.2 RESULTS AND DISCUSSION

As in the previous study, the number of data points was reduced by randomly selecting 65,000 cosine values per text using a simple random sampling technique. An ANOVA comparing the idiolects showed a significant difference between the four authors (*F*(1, 1040000) = 2650.68, *p* < .001). Contrary to what was predicted differences were also found between the texts for each of the authors (Eliot: (*F*(3, 260000) = 1305.17, *p* < .001; *Mean Cosine* = .034, *SD* = .07; Dickens: (*F*(3, 260000) = 645.62, *p* < .001; *Mean Cosine* = .020, *SD* = .06; Woolf: (*F*(3, 260000) = 167.80, *p* < .001; *Mean Cosine* = .019, *SD* = .045; Joyce: (*F*(3, 260000) = 2899.20, *p* < .001; *Mean Cosine* = .021, *SD* = .073). This again suggests no support for the idiolect hypothesis. In the LSA comparison between the texts of one author described in the previous study, most homogeneity was found in the texts by Eliot. This effect was replicated in the internal homogeneity analysis, suggested by the highest LSA cosine values.

For the sociolect-gender hypothesis a significant effect was found between gender (*F*(1, 1040000) = 1699.02, *p* < .001), with texts written by female authors having higher cosine values (*Mean Cosine* = .026, *SD* = .058) than those written by male authors (*Mean Cosine* = .020, *SD* = .068). In addition, differences were found for within-gender texts (female: *F*(1, 520000) = 1929.72, *p* < .001; male: *F*(1, 520000) = 1660.43, *p* < .001).

As for the sociolect-period hypothesis, differences were found between periods (*F*(1, 1040000) = 2563.28, *p* < .001), but also between the authors within a period (Realist: *F*(1, 520000) = 4788.1, *p* < .001; Modernist: *F*(1, 520000) = 61.08, *p* < .001). In the previous analysis we saw that Realist texts share more semantic concepts between them. Similarly, for the semantics between parts of the Realist texts cosine values are higher than for Modernist texts (Realist: *Mean Cosine* = .027, *SD* = .067; Modernist: *Mean Cosine* = .020, *SD* = .060).

An explanation for the results we have found in the previous studies might indeed lie in the internal semantic homogeneity. This analysis replicated the finding in study 3 that Modernist texts seem to be more diverse than Realist texts. This is an important finding for corpus linguistic analyses of modern literary texts in general, but also for the validity of the Modernist-code hypothesis. If it is true that Modernist authors experiment more with their literary products (see Fokkema and Ibsch, 1988), then it is still possible to keep up a Modernist hypothesis: Certain semantic fields might still be more prominent in these texts. However, their overall frequency is low because Modernist texts miss the homogeneity Realist texts have.

7. Conclusion

We tested hypotheses initially brought forward by in Fokkema and Ibsch (1988), who argued that selected authors use selected semantic fields. The word frequency of the contents of these fields would predict frequency patterns in idiolect and literary period. We tested idiolect, sociolect-gender, sociolect-time and Modernist-code hypotheses derived from this study using Boolean models and vector models. A total of 16 literary texts were used balanced across author (Eliot, Dickens, Woolf, Joyce), gender (female, male) and literary period (Realism, Modernism). Two models were used to test these hypotheses, a binary Boolean model and a scaling vector model. Both methods are very common the field of corpus linguistics (see Louwerse and Van Peer, 2003 for an overview).

Initial Boolean analyses suggested no evidence for any one of the four hypotheses, possibly because of the semantic fields that were selected and the Boolean method that was used. Results were replicated in a vector model using the semantic fields. A vector analysis comparing the general content between the groups of texts and comparing the various parts within each text, showed that the semantic homogeneity in literary texts is an important confounding variable. Because of this, drawing conclusions from a literary text as a whole, rather than its parts might be problematic. A vector model can partly solve this problem, by taking into account every part of the text. But drawing conclusions from semantic similarities within an author can be equally problematic, because authors tend to change their style and semantic space between texts. Similarly, semantic similarities within a literary period are difficult to determine because of the overall variations. As pointed out in the beginning of this study, the lack of internal homogeneity in one text, between texts and between authors can be explained by the (semantic) deviation from the norm the author tries to establish. These variations are exactly what makes the idiolect and sociolect of literary texts unique, and is in fact what makes those texts literary.

Acknowledgements

This research was partially supported by the National Science Foundation (SES 9977969, REC 0106965, REC 0113669) and the Institute of Education Sciences (IES) (R 305G020018). 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 funding agencies.

Notes

[1] Whereas the Modernist-code hypothesis, sociolect-time and idiolect hypotheses are directly derived from Fokkema and Ibsch (1988), the sociolect-gender hypothesis is not. However, given the theory of a group code, a sociolect-gender hypothesis seems justified.

[2] *Portrait* is used as an abbreviation for *Portrait of the Artist as a Young Man*.

References

Baeza-Yates, R. and Ribeiro-Neto, B. (Eds.) (1999). *Modern Information Retrieval*. ACM Press, New York. 513 p.

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge University Press, Cambridge, UK. 315 p.

Eco, U. (1977). *A Theory of Semiotics*. Indiana University Press, Bloomington. 368 p.

Fellbaum, C. (Ed.) (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA. 500 p.

Fokkema, D. and Ibsch, E. (1987) *Modernist Conjectures. A Mainstream in European Literature 1910-1940*. Hurst, London. 330 p.

Foltz, P. W., Kintsch, W. and Landauer, T. K. (1998). The Measurement of Textual Coherence With Latent Semantic Analysis. *Discourse Processes, 25,* pp. 285-307.

Graesser, A., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., Person, N., and the Tutoring Research Group. (2000). Using Latent Semantic Analysis to Evaluate the Contributions of Students in Autotutor. *Interactive Learning Environments, 8,* pp. 149-169.

Hu, X., Cai, Z., Franceschetti, D., Penumatsa, P., Graesser, A. C., Louwerse, M. M., McNamara, D. S. and the Tutoring Research Group (2003). *LSA: The First Dimension and Dimensional Weighting*. *Proceedings of the 25rd Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.

Jakobson, R. (1987). Linguistics and Poetics. In R. Jakobson, *Language in Literature*. Harvard University Press, Cambridge, MA, pp. 62-94.

Kintsch, W.(2000). Metaphor Comprehension: A Computational Theory. *Psyhonomic Bulletin and Review, 7,* pp. 257-266

Landauer, T. K., and Dumais, S. T. (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of the Acquisition, Induction, and Representation of Knowledge. *Psychological Review, 104,* pp. 211-240.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes, 25,* pp. 259-284.

Lotman, J. (1977). *The Structure of the Artistic Text*. University of Michigan, Ann Arbor. 300 p.

Louwerse, M.M. and Van Peer, W. (eds.) (2002). *Thematics: Interdisciplinary Studies.* John Benjamins, Amsterdam/Philadelphia. 430 p.

Martindale, C. (1990). *The Clockwork Muse*. Basic Books, New York. 411 p.

Pennebaker, J. W. (2002). What our Words Can Say about Us: Towards a Broader Language Psychology. *Psychological Science Agenda*, *15*, 8-9.

Project Gutenberg, http://www.ibiblio.org/gutenberg

Sebeok, T. A. (1991). *A Sign is Just a Sign*. Indiana University Press, Bloomington. 178 p.

The Online Books Page, http://onlinebooks.library.upenn.edu

The Oxford Text Archive, http://ota.ahds.ac.uk

Wardhaugh, R. (1998). *An Introduction to Sociolinguistics*. Blackwell, Oxford, UK. 464 p.

Watson, G. (1994). A multidimensional analysis of style in Mudrooroo Nyoongah’s prose works. *Text*, *14*, pp. 239-285.

Wellek, R. and Warren, A. (1963). *Theory of Literature*. Cape, London. 382 p.