CHAPTER 12

Computationally discriminating literary from non-literary texts

Max Louwerse, Nick Benesh & Bin Zhang

Three computational linguistic methods are presented to discriminate literary from non-literary texts. In the first study, a hierarchical clustering technique of results obtained from Latent Semantic Analysis showed a clustering of literary versus non-literary texts. The second study used the frequencies of shared bigrams across the text, resulting in a 100% correct classification of literary versus non-literary texts. The third study used unigrams yielding a 94% correct classification into literary versus non-literary texts. The final two studies using a larger sample of texts showed that the high classification performance cannot be attributed to specific texts. These findings provide evidence that distinguishing literature from non-literature can be done with high accuracy and with relatively simple computational linguistic techniques.

Keywords: computational linguistics, stylistics, genre, bigram analysis, latent semantic analysis, classification techniques

1. Introduction

It does not take much effort for those who visited him to recall his office on Muntstraat 4 in Utrecht, on the second floor immediately left from the squeaky stairs. Coffee in a plastic cup from the machine at the end of the corridor; inside his office two desks covered with piles of papers, the ones that had not quite made it into the binders that filled the shelves on the wall. Radiators turned high, the door slightly open as a silent invitation to researchers in the field of empirical studies of literature. At the time, the first author was a student in Literary Studies and research assistant of Willie van Peer.

Literary studies students study literature. If somebody asks what constitutes as literature one ought to reply with terms like "aesthetic", "deconstruction", "fictionality", "defamiliarization", "*syuzhet*", "foregrounding" and "literariness" (Jefferson & Robey 1986). The first author recited these same terms when his research assistants, the

co-authors of this paper, asked that question. The problem with this answer is that it may be possible to identify prototypical examples of literary language from literature (Fabb 2002; Fowler 1996), but whether literary texts overall are linguistically different from non-literary texts is a question that has not been satisfactorily answered. In fact, very little is known about the language that distinguishes literature from non-literature and whether computational linguistic techniques can offer answers (see Louw 1993; Sinclair 2004; Stubbs 2005).

Perhaps the answers these computational linguistic techniques provide are "utterly naïve" (van Peer 1989: 302) because the fundamental literary ingredients are reduced to lower levels of linguistic organization. That is, "no level of (mathematical) sophistication is able to overcome the problem that the processes of meaning constitution have been eliminated before the analysis is undertaken" (van Peer 1989: 302). Indeed, the text and its meaning should be considered (Sinclair 2004) and one should refrain from taking a text to pieces (Sinclair 1966).

The aim of the current chapter was to do exactly that: use computational linguistic techniques that take literary and non-literary texts to pieces to provide, what van Peer (1989: 302) called, "utterly naïve" answers. More specifically, this paper asks the question of whether distinctions can be made between literary from non-literary texts by using three computational linguistic methods: 1) a higher-order co-occurrence analysis of the words in all texts; 2) a computation across all texts by considering the frequencies of bigrams, or word pairs, 3) an account per text of the frequencies of specific words or types of words. The findings presented in these five studies provide evidence that distinguishing literature from non-literature can be done with high accuracy and with relatively simple computational linguistic techniques.

2. Study 1

One way to investigate the semantic content of texts is by looking at the semantic neighbors of words. Sentences like "The researcher worked on foregrounding in literary texts", "The researcher worked on stylistics in literary texts", and "The researcher worked on the empirical studies of literary texts" suggest that "foregrounding", "stylistics" and "empirical studies" have semantic content in common. Sentences or paragraphs however often do not have the same semantic context, resulting in a sparsity problem. Higher-order relationships between words can then be the solution, by considering the neighbors of the neighbors of the neighbors etc. of words. Latent Semantic Analysis (LSA) is a statistical technique that estimates the semantic relations between words, sentences, paragraphs or texts by

considering these higher-order co-occurrences of words. Meaning is captured by mapping initially meaningless words into a continuous high-dimensional semantic space, which more or less simulates cognition (Landauer 2002). More specifically, a first-order process associates stimuli (words) and the contexts they occur in (documents). Stimuli are paired based on their contiguity or co-occurrence. These local associations are next transformed by means of Singular Value Decomposition into a small number of dimensions (typically 300) yielding more unified knowledge representations by removing noise. Because it considers associations between concepts, represented by words, LSA could be seen as a theory of knowledge representation, induction and language acquisition (Landauer & Dumais 1997; Landauer, McNamara, Dennis & Kintsch 2007; for a comprehensive introduction to LSA, see Kintsch 2002; see also Louwerse 2004, Louwerse & van Peer 2006; in press, for a demonstration of how LSA can be extremely helpful in the analysis of literary texts).

2.1 Materials

In this first study, we compared the semantic content in literary and non-literary texts using LSA in order to determine whether semantic content allows for clustering of texts in at least these two groups. The materials used were nine literary and seven non-literary texts. Texts were considered literary if they appeared in Zane's book (2007) *Top Ten List*, whereby he polled 125 British and American top authors asking them to pick the top 10 books of literature of all time. Nine of the top ten books were made electronically available; one could not (Marcel Proust's *In Search of Lost Time*). Non-literary texts were chosen from a set of corpora that was used in previous studies (Crossley & Louwerse 2007). An overview of these 16 texts is presented in Table 1. Note that numbers of texts in the two conditions slightly differed, but that analyses in this study and further studies are not sensitive to unequal cell sizes.

The nine literary texts differed in size (e.g., *War and Peace* and *The Great Gatsby*), genre (e.g., *Letters of Anton Chekhov*, *Hamlet*, *Anna Karenina*), original language versus English translations (e.g., *Madame Bovary*, *Middlemarch*) and year of publication (e.g., *Hamlet*, *Lolita*).

The seven non-literary texts differ in formality (e.g., *Santa Barbara Corpus*, *Trains*), in written versus spoken register (e.g., *Wall Street Journal*, *Edinburgh Map-Task Corpus*) and modality (e.g., *Switchboard*, *Santa Barbara Corpus*, *New York Times*). The Edinburgh MapTask Corpus (Human Communication Research Centre, 1993) is a task-based corpus that is the linguistic product of a cooperative task involving two participants. Instruction givers have a marked route on

Table 1. Texts and corpora used in Studies 1–3

Author	Title	Year	Word count	Genre
L. Tolstoy	Anna Karenina	1877	358,689	Literature
W. Shakespeare	Hamlet	1601	32,756	Literature
A. Chekhov	Letters of Anton Chekhov	1888	118,639	Literature
V. Nabokov	Lolita	1955	116,752	Literature
G. Flaubert	Madame Bovary	1856	117,554	Literature
G. Eliot	Middlemarch	1871	326,025	Literature
M. Twain	Adventures of Huckleberry Finn	1884	116,547	Literature
F. Fitzgerald	The Great Gatsby	1925	50,110	Literature
L. Tolstoy	War and Peace	1869	572,535	Literature
	Memphis MapTask	2008	343,050	Dialogue (task-based)
	Edinburgh MapTask	1991	2,731,809	Dialogue (task-based)
	New York Times	1996	29,560,931	Newspaper
	Santa Barbara	2000	464,932	Dialogue (informal)
	Switchboard	1997	3,702,166	Dialogue (telephone)
	TRAINS	1995	84,149	Dialogue (text-to-text)
	Wall Street Journal	1996	17,956,625	Newspaper

their map and give directions to the instruction followers who have no route. The maps are not identical, which elicits unscripted problem solving dialog. The Memphis Multimodal MapTask Corpus (Memphis MapTask Corpus; Louwerse, Bard, Steedman & Graesser 2004) is similar to the Edinburgh Corpus except that different maps were chosen and the setup focused on multimodal communication. The corpus was selected here as a comparison to the Edinburgh MapTask Corpus, with a subset of only 32 (out of the 256 dialogues) being used. The TRAINS Corpus (Allen & Heeman 1995) is based on the routing and scheduling of freight trains. The corpus shares with MapTask its basis as a task-based corpus, but it is more temporal and directional in nature than the spatial MapTask Corpus. Two non-task-based dialogue corpora are the Switchboard Corpus (Godfrey & Holliman 1993) and the Santa Barbara Corpus (Du Bois, Chafe, Meyer & Thompson 2000). The Switchboard Corpus is a collection of 2,400 two-sided random topic telephone conversations taken from 543 speakers from all areas of the United States. The Santa Barbara Corpus is a collection of natural speech recordings taken from people across the United States. Finally, the New York Times and Wall Street Journal corpora consist of all articles from those newspapers from 1996. The texts included in this

study are extremely diverse. Because of the diversity of the literary as well as the non-literary texts an accurate classification of these texts into literary and non-literary on the basis of semantic content is far from obvious.

2.2 Results

For LSA analyses, a 'knowledge base' is needed in the form of an LSA space. The semantic relation between two or more words, sentences, paragraphs, or – in the current case – texts can be computed using this knowledge base. The general Touchstone Applied Science Associates (TASA) Corpus is commonly used to create such a space and is only used for that purpose in this study. The TASA Corpus consists of approximately 10 million words of unmarked high school level English texts on Language Arts, Health, Home Economics, Industrial Arts, Science, Social Studies, and Business. This corpus is divided into 37,600 documents, averaging 166 words per document, and is considered one of the benchmark corpora in computational linguistics, because it approximates the language familiarity of a college level student (Landauer & Dumais 1997). The matrix of 16x16 cosine values representing the semantic similarities between the texts was submitted to a hierarchical clustering analysis, using a between-groups linkage clustering method and a squared Euclidean distance measure. The dendogram resulting from this analysis is presented in Figure 1.

Figure 1 shows that by taking the semantic content of the texts, LSA is able to discriminate between two groups of texts, literary and non-literary. Within the non-literary group, the task-oriented dialogues cluster together (*Edinburgh MapTask Corpus* and *Memphis MapTask Corpus*, as well as *Trains*), followed by a grouping of the other dialogues. Next, the two newspaper corpora clustered together. The selected dialogues of the two *Maptask* corpora considerably differed in size (the selected Edinburgh Corpus being almost eight times larger). The fact

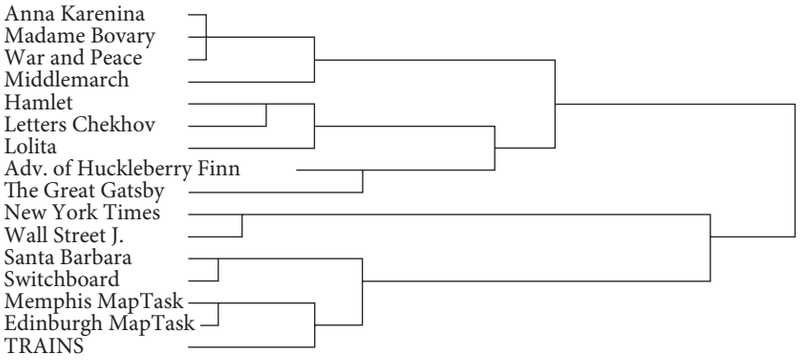

Figure 1. Hierarchical clustering of LSA cosine values between all 16 texts.

that the two clustered closely together therefore suggests that LSA is not sensitive to corpus size. Zane's (2007) top 3 literary works (*Anna Karenina*, *Madame Bovary*, *War and Peace*) clustered together, with *Middlemarch* being added to this cluster. Another cluster that emerges is that of two novels that can be considered "the canon of American self-images and values" (Rowe 1988: 13) (*Adventures of Huckleberry Finn*, *The Great Gatsby*). Those texts with a different style, because of genre (*Letters of Anton Chekhov*, *Hamlet*) or the narrating style (*Lolita*; see Green 1987) also formed a cluster. The main result of this analysis, however, is that an analysis based on semantic relations across all lexical items of the texts allows for a categorization of literary and non-literary texts.

Identical results as those in the hierarchical clustering method were obtained using an ALSCAL Multidimensional Scaling (MDS) representation algorithm, with the matrix of LSA cosine values being transformed into a matrix of Squared Euclidean distances. MDS does not result in a hierarchical clustering, but the stimulus coordinates, followed an identical pattern as the hierarchical clustering algorithm. The advantage of these stimulus coordinates is that they can later be used for correlational analyses. The MDS fitting of the data was satisfactory (*Kruskal's stress 1* = .25; R^2 = .84) with a one-dimensional scaling.

The advantage of an LSA analysis is that all semantic content is taken into account when clustering the text. The drawback is that because a higher-order co-occurrence technique is used, it is difficult to determine what linguistic information is accountable for these clusters. As discussed in Crossley & Louwerse (2007) and Louwerse, Lewis and Wu (in press), a bigram analysis can account for this problem. The same 16 texts were classified using a bigram analysis in Study 2.

3. Study 2

Bigrams are combinations of two words occurring in a corpus. A text consisting of a sentence "John loves Mary and Mary John" consists of the bigrams "John loves", "loves Mary", "Mary and", "and Mary", "Mary John". One could take the frequency of these bigrams and compare them with the frequencies of bigrams shared in other texts. If the bigram "and Mary" is high in Text 1 and 2, and significantly lower in Text 3 and 4, two groups of texts can be identified.

Bigrams (or rather n-grams) have been used in a number of language models such as determining the probability of a sequence of words, speech recognition models, spelling correction, machine translation systems, and optical character recognizers (Jurafsky & Martin 2000; Manning & Schutze 1999). For instance, Crossley & Louwerse (2007) used bigrams for register classification of spoken and

written corpora, and Louwerse, Lewis & Wu (in press) used bigrams to categorize the genres in Shakespeare's plays. Study 2 tested whether bigrams also allow for a classification of texts in literature versus non-literature.

3.1 Materials

In the second study, we used the same seven corpora and nine texts as in Study 1.

3.2 Results

The frequency of all bigrams in each of the seven corpora and nine texts was computed and normalized to account for corpus size. Next, only bigrams that were shared across all materials were selected, resulting in a total of 61 bigrams. The normalized frequencies of these bigrams were compared between the literary and the non-literary texts using a Mann-Whitney test. Those bigrams that yielded significant differences ($p < .05$) are presented in Table 2. All significant Z-values showed negative scores, with a higher occurrence of these bigrams in literary than in non-literary texts.

Next, those bigrams that yielded a significant difference between literary and non-literary texts at $p < .01$. This selection included 10 bigrams that were submitted

Table 2. Mann-Whitney test between bigram frequencies of literature and non-literature

Bigram	*U*	*Z*	Bigram	*U*	*Z*
and in	0	−3.33**	is that	2	−3.12**
and so	10	−2.28*	it to	14	−1.85*
and the	11	−2.17*	it was	10	−2.28*
and what	5	−2.81*	me to	11	−2.17*
and with	3	−3.02**	not the	15	−1.75*
as I	5	−2.81*	of a	4	−2.91**
as it	1	−3.23**	of this	5	−2.81*
but the	6	−2.70*	on a	14	−1.85*
by the	10	−2.28*	on the	14	−1.85*
for a	6	−2.70*	out of	4	−2.91**
I am	5	−2.81*	that I	11	−2.17*
I had	9	−2.38*	to be	12	−2.06*
I will	6	−2.70*	to see	2	−3.12**
in a	6	−2.70*	was a	4	−2.91**
in my	6	−2.70*	what is	11	−2.17*
in the	10	−2.28*	with a	3	−3.02**
into the	2	−3.12**	with the	11	−2.17*

* $p < .05$, ** $p < .01$.

to a discriminant analysis. Three bigrams classified literary from non-literary text with 100% accuracy. These bigrams and their discriminant function coefficients are presented in Table 3, the classification results in Table 4 and the Z-scores in Table 5.

Table 3. Discriminant function coefficients per bigram

Bigram	Discriminant function coefficient
and in	75.692
and with	78.845
was a	38.229

Table 4. Classification results bigrams

		Predicted	
		Literature	Non-literature
Original	Literature	100% (9)	0 (0)
	Non-literature	0 (0)	100% (7)

Note: Numbers in parentheses are actual counts. Percentage of correctly classified, 100, Eigenvalue, 16.71, Wilk's λ, .056, χ^2, 35.93, $p < .001$.

Table 5. Discriminant scores per text

Text	*Z*
Anna Karenina	4.29
The Great Gatsby	4.26
War and Peace	3.64
Madame Bovary	3.53
Hamlet	3.49
Middlemarch	3.06
Adventures of Huckleberry Finn	2.96
Letters of Anton Chekhov	2.95
Lolita	2.16
Switchboard	−2.44
New York Times	−2.93
Wall Street Journal	−4.15
Santa Barbara	−4.21
Memphis MapTask	−5.02
Edinburgh MapTask	−5.77
TRAINS	−5.83

The scores per text showed a clear distinction between the literary (positive scores) and non-literary (negative scores) texts. To determine how the results from Study 1 were related to those of Study 2, the stimulus coordinates from the MDS results in Study 1 were taken and compared with the discriminant scores of study 2. A strong correlation was found ($r(16) = .91$, $p < .001$). This shows that both the LSA and bigram analysis yielded very similar results, despite the fact that LSA emphasized semantic content expressed by lexical items, whereas shared bigrams typically highlight functional items (see also Louwerse, Lewis & Wu, in press).

As Crossley and Louwerse (2007) have argued, one of the advantages of a bigram analysis over an LSA analysis is that it allows for getting an insight in the linguistic patterns responsible for differences. The three bigrams variables in the current discriminant analysis suggest that literary texts are typically written in a past tense ("was a"). Moreover, literary texts had a higher frequency of preposed adverbial phrases as indicated by bigrams like "and in", and "and with". This may indicate a more frequent use of thematic reorientation, as well as cognitive reorientation, in literary texts compared to non-literary texts (Givón 1993). That conclusion may be premature based on these three variables, but the fact is that a perfect classification was obtained using the frequencies of only these bigrams, very similar to that using the semantic content of the texts.

Both Studies 1 and 2 used data mining techniques: on the basis of frequently occurring patterns, texts were classified in literary and non-literary. Can the findings obtained from LSA and bigram techniques be extended to unigrams? And do variables reported in previously published work in literary studies yield a similar classification of the current corpora and texts? These questions are answered in Study 3.

4. Study 3

In the third study, we followed up on a previous one by van Peer (1986). In that, linguistic features in quality literature are compared with those in popular literature, as defined according to their production process and distribution channels. Van Peer (1986) argued that quality literature has a higher number of occurrences of 3rd person narrators, but a lower number of 1st person narrators than popular literature. Moreover, quality literature has a higher number of 1st names and last names than popular literature. Study 3 investigated the frequency of names and narrators in the seven corpora and nine texts.

4.1 Materials

Study 3 used the same nine literary texts and seven non-literary corpora as in the previous two studies.

4.2 Results

For each text, the frequency of 1st person narrators, 3rd person narrators, first names and family names were computed. First person narrators were operationally defined, following van Peer (1986) using the pronouns "I", "my", "mine", third person narrators by the pronouns "he", "she", "his", "her", "him", "hers". Clearly, this is an operational definition, since pronouns could refer to narrators as well as to characters. First names were identified using 4,275 female first names and 1,219 male first names as well as 88,799 last names from the U.S. Census Bureau data (1994). Name and pronoun frequencies were normalized to account for text size.

As in the previous study, a Mann-Whitney test was conducted to determine which of these variables yielded a significant difference between literary and non-literary texts. The results are presented in Table 6, showing that only pronouns related to the 3rd person narrator yielded a difference.

Next, the 3rd person narrator variable was entered in a discriminant analysis. A total of 93.8% of the cases was correctly classified only using the third person pronoun (discriminant coefficient .664). Classification results are presented in Table 7, scores per text in Table 8.

The one text that was misclassified is *Letters of Anton Chekhov*. *Adventures of Huckleberry Finn, Hamlet* and *Lolita* were classified correctly, but had similarly

Table 6. Mann-Whitney test

Variable	*U*	*Z*
First name	17	−1.53
Last name	30	−0.16
1st person narrator	14	−1.85†
3rd person narrator	0	−3.33**

Note: $^{**}p < .01$, $\dagger p < .07$. Negative *Z* scores denote a higher frequency in literary texts.

Table 7. Classification results third person narrator

		Predicted	
		Literature	**Non-literature**
Original	Literature	88.9 (8)	11.1 (1)
	Non-literature	0 (0)	100.0 (7)

Note: Numbers in parentheses are actual counts. Percentage of correctly classified, 93.8, Eigenvalue, 1.97, Wilk's λ, .337, χ^2, 14.695, $p < .001$.

Table 8. Discriminant scores per text

Text	Discriminant score
Anna Karenina	2.91
Madame Bovary	2.88
War and Peace	1.73
Middlemarch	1.66
The Great Gatsby	1.39
Adventures of Huckleberry Finn	0.42
Lolita	0.31
Hamlet	−0.09
Letters of Anton Chekhov	−0.80
New York Times	−1.04
Santa Barbara Corpus	−1.16
Switchboard	−1.38
Wall Street Journal	−1.38
Memphis MapTask	−1.77
Edinburgh MapTask	−1.85
TRAINS	−1.85

low discriminant scores. It is noteworthy that two of these texts (*Letters of Anton Chekhov, Lolita*) also scored low in the classification in Study 2, and all three of these texts clustered together in Study 1. In fact, a strong correlation was found between the discriminant scores obtained from the texts in Study 2 and 3 ($r(16) = .852$, $p < .001$).

The main finding of this study is that a near-perfect classification between literary and non-literary texts was obtained and that this classification was similar to that of Study 1 and Study 2.

5. Study 4

Studies 1–3 showed that different computational linguistic techniques (LSA, bigrams, unigrams) obtained very similar results for the 16 texts. Perhaps the similarity can be explained by the texts that were used. Indeed, the argument could be made that Studies 1–3 primarily compared literary texts and spoken dialogue. This argument does not quite hold because newspaper articles were included, but the majority of the materials were dialogues. To account for issues regarding generalizability of the selected materials, a fourth study used the variables from Studies 2 and 3, the three bigrams ("I was", "and in" "and with") and the 3rd person narrator, and ran a discriminant analysis on a different set of texts.

5.1 Materials

Texts included 119 literary texts, as identified by Zane (2007), without the top 10 literary texts used in the previous three studies. These included *A Midsummer's Night Dream*, *A Portrait of the Artist as a Young Man*, poems of Emily Dickinson, *The Ancient Mariner*, and *The Divine Comedy*. Non-literary texts included 55 random texts from Project Gutenberg (Hart 2004) that were not classified by the Library of Congress Classification (LCC) System as Language and Literatures (A-Z, with the exception of PN-PZ). These included genres identified by LCC as General Works; Philosophy, Psychology, and Religion; Auxiliary Sciences of History; General and Old World History; History of America; Geography, Anthropology, and Recreation; Social Sciences; Political Science; Education, etc., with titles of texts like *The Outline of Science*, *A Text-Book of the History of Painting*, *Across Unknown South America*, *Custom and Myth*, and *History of the United States*. As in the previous studies, both the literary and non-literary texts differed in content and genre.

5.2 Results

Normalized frequencies were computed for the three bigrams and the third person narrator variables for each of the 174 texts. These frequencies were next entered in a discriminant analysis comparing the literary with the non-literary texts. If the results from the previous studies permit being generalized, these four variables should allow for a higher-than-chance classification into literature and non-literature. Results again showed that bigrams and third person pronouns allow for a distinction between literary and non-literary texts, with 87.4% of the texts correctly classified. Two of the bigrams and the third person narrator contributed to this classification (Table 9). Classification results are presented in Table 10.

These results show that the findings obtained from Studies 1–3 cannot be attributed to specific texts. When a larger and different set of texts is used, bigrams and third person narrators allow for an equally high classification performance of literary versus non-literary texts.

Table 9. Discriminant function coefficients

Variables	Function coefficient
and in	−.472
and with	.621
3rd person narrator	.922

Note: Negative scores denote a higher frequency in literary texts, positive scores a higher frequency in non-literary texts.

Table 10. Classification results bigram and third person narrator 174 texts

		Predicted	
		Literature	**Non-literature**
Original	Literature	89.1% (106)	10.9% (13)
	Non-literature	16.4% (9)	83.6% (46)

Note: Numbers in parentheses are actual counts. Percentage of correctly classified, 87.4, Eigenvalue, 1.481, Wilk's λ, .486, χ^2, 154.921, $p < .001$.

One could argue that in the Studies 1–4 cell sizes are different. For instance, in Study 4, there are far more literary texts than non-literary texts. As stated earlier, discriminant analysis is relatively insensitive to groups not being of equal size. To illustrate this, the 55 non-literary and a random sample of 55 literary texts were entered in the discriminant analysis. Classification performance was the same, with 85.5% of the cases correctly classified.

To extend the generalizability of the findings reported so far, a final analysis was conducted in Study 5.

6. Study 5

Study 1 showed that LSA allows for a clustering of the corpora and texts into literary and non-literary, Studies 2 and 3 showed that the same classification can be obtained using shared bigrams and even unigrams. Study 4 extended these findings by considering a different set of literary and non-literary texts, thereby ruling out the possibility that the results obtained in Studies 1–3 are due to genre (dialogue and newspaper versus novels) rather than literary versus non-literary differences. However, the question can be raised whether these studies compare literature with non-literature. Perhaps the classification results can be explained by a comparison of narrative (literary texts) and non-narratives (dialogues, newspaper articles, histories, textbooks). To answer these questions Study 5 further investigated the literary aspect of the texts by comparing two groups of narratives, what van Peer's (1986) called 'quality literature' and 'popular literature'.

6.1 Materials

The same 119 quality literature texts from Study 4 were used. As a popular literature comparison, 42 Star Wars novels (1979–2003) were made electronically available. These were part of form a collection of novels approved by George Lucas and were contiguous with the original three Star Wars movies. Much like a long

running Star Wars TV series with new episodes taking into account previous ones, the novels revolve around the main characters' adventures to save the galaxy from various catastrophes. The authors of the novels vary in many dimensions, such as notoriety, influence on the overall Star Wars time line, and general experience in the science fiction genre.

6.2 Results

Of the four variables (three bigrams and third person narrator) entered in the discriminant analysis only the bigram "and in" was kept in the model (discriminant function coefficient = 50.517). This variable alone classified 86.3% of the cases correctly. Classification results are given in Table 11.

Five out of the 119 literary texts were misclassified on the basis of the frequency of the bigram "and in". These were *The Importance of Being Earnest*, *Uncle Vanya*, *To Kill a Mockingbird*, poems of Emily Dickinson and *Hedda Gabbler*. The reason for this misclassification is not clear. Though three of these five texts are not novels but plays and poems, the remaining 114 literary texts contain many plays and poems that were classified correctly, such as *King Lear*, *Henry IV*, *Henry V*, *The Aeneid*, *The Iliad* and *The Ancient Mariner*.

Study 5 showed that when distinguishing quality from popular literature, a high classification performance can be obtained using the normalized frequency of one single bigram (86.3%). Moreover, the classification results show that the quality literary texts are unlikely to be classified as popular literature (95.8%), but popular literature may get classified as quality literature (40.5%). That is, what is quality literature remains quality literature, but in some cases what is popular literature could be considered quality literature.

7. Conclusion

According to Fowler (1996: 21) "Literature is a creative use of language". The five studies presented in this paper show that different computational linguistic

Table 11. Classification results quality literature and popular literature

		Predicted	
		Literature	**Star Wars**
Original	Literature	95.8 (114)	4.2 (5)
	Star Wars	40.5 (17)	59.5 (25)

Note: Numbers in parentheses are actual counts. Percentage of correctly classified, 86.3, Eigenvalue, .448, Wilk's λ, .691, χ^2, 58.677, $p < .001$.

techniques allow for a high performance classification into literary and non-literary texts on the basis of the language in literary and non-literary texts. Study 1 showed that higher-order occurrence techniques like LSA can cluster corpora and texts, Study 2 showed similar results can be obtained using three shared bigrams, and Study 3 showed even pronoun frequencies allow for an accurate classification. Moreover, despite the different techniques used in Study 2 and 3, a correlation of discriminant scores was obtained. Study 4 showed that this classification performance cannot be attributed to the selection of texts. On the contrary, when a larger set of corpora and texts of literary and non-literary variety were used, classification performance remains high. In fact, even when two sets of narratives (quality literature and popular literature) were compared, classification performance remains considerably higher than chance (86.3%). Moreover, what Study 5 showed is that quality literature was unlikely to be classified as popular literature, but what is popular literature might be classified as quality literature.

The finding that different computational linguistic techniques yield similar results is important for future analyses. For instance, both a focus on lexical items (content words) and a focus on functional items (pronouns, conjunctions) yielded similar results (see also Louwerse, Lewis & Wu , in press). An overview of some of the differences is presented in Table 12.

The five studies reported in this paper served a number of purposes. First, they presented some straightforward computational linguistic techniques like Latent Semantic Analysis, shared bigram frequencies and word frequencies available to researchers investigating linguistic patterns in text and discourse. Secondly, the studies reported here shed light on the question whether literary texts can be distinguished from non-literary texts using language features, not in a selected text sample but throughout the text. Twenty years ago, if somebody asked in his office on the second floor immediately left from the squeaky stairs what constituted literature, the reply would have been with terms like "aesthetic", "deconstruction", "fictionality", "defamiliarization", "*syuzhet*", "foregrounding" and "literariness". Twenty years later, we can reply with the bigram "and in", an empirically valid answer. Whether

Table 12. Differences between computational linguistic techniques

	Word count	Shared bigrams	LSA
Sensitivity to paragraph size	–	–	+
Sensitivity to text size	–	–	–
Variables user defined	+	–	–
Sensitivity to exact string match	+	+	–
Sensitivity to lexical items	+	–	+
Sensitivity to grammatical items	+	+	–
Sensitivity to linguistic patterns	–	+	–

that computational linguistic answer is utterly naïve, as van Peer (1989) suggested, we leave open for discussion.

8. Acknowledgments

This research was supported by grant NSF-IIS-0416128. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding institution.

References

Allen, J. & Heeman, P.A. 1995. *TRAINS Spoken Dialog Corpus* [CD-ROM]. Philadelphia PA: Linguistic Data Consortium.

Crossley, S.A. & Louwerse, M.M. 2007. Multi-dimensional register classification using collocations. *International Journal of Corpus Linguistics* 12: 453–478.

Du Bois, J.W., Chafe, W.L., Meyer, C. & Thompson, S.A. 2000. *Santa Barbara Corpus of Spoken American English* [CD-ROM]. Philadelphia PA: Linguistic Data Consortium.

Fabb, N. 2002. *Language and Literary Structure: The Linguistic Analysis of Form in Verse and Narrative*. Cambridge: CUP.

Fowler, R. 1996. *Linguistic Criticism*. Oxford: OUP.

Givón, T. 1993. *English Grammar: A Function-based Approach*. Amsterdam: John Benjamins.

Godfrey, J.J. & Holliman, E. 1993. *Switchboard-1* [CD-ROM]. Philadelphia PA: Linguistic Data Consortium.

Graff, D. 1995. *North American News Text Corpus* [CD-ROM]. Philadelphia PA: Linguistic Data Consortium.

Green, M. 1987. Tolstoy and Nabokov. The morality of Lolita. In *Nabokov's Lolita*, H. Bloom (ed.), 13–33. New York NY: Chelsea House.

Hart, M. 2004. *Project Gutenberg*. Retrieved November 20, 2007, from http://www.gutenberg.org/

Human Communication Research Centre. 1993. *HCRC Map Task Corpus* [CD-ROM]. Philadelphia PA: Linguistic Data Consortium.

Jurafsky, D. & Martin, J.H. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River NJ: Prentice-Hall.

Jefferson, A. & Robey, D. 1986. *Modern Literary Theory: A Comparative Introduction*. London: Batsford.

Kintsch, W. 2002. On the notions of theme and topic in psychological process models of text comprehension. In *Thematics: Interdisciplinary Studies*, M.M. Louwerse & W. van Peer (eds), 157–170. Amsterdam: John Benjamins.

Landauer, T.K. & 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: 211–240.

Landauer, T.K., McNamara, D.S., Dennis, S. & Kintsch, W. (eds). 2007. *Handbook of Latent Semantic Analysis*. Mahwah NJ: Lawrence Erlbaum.

Landauer, T.K. 2002. On the computational basis of learning & cognition: Arguments from LSA. *The Psychology of Learning & Motivation* 41: 43–84.

Louw, B. 1993. Irony in the text or insincerity in the writer? In *Text and Technology. In honour of John Sinclair*, M. Baker, G. Francis & E. Tognini-Bonelli (eds), 157–176. Amsterdam: John Benjamins.

Louwerse, M.M., Bard, E., Steedman. M. & Graesser, A.C. 2004. *Memphis Multimodal MapTask Corpus* [DVD]. Memphis TN: University of Memphis.

Louwerse, M.M., Lewis, G. & Wu, J. In press. Unigrams, bigrams and LSA: Corpus linguistic explorations of genres in Shakespeare's plays. In *New Directions in Literary Studies*, W. van Peer & J. Auracher (eds). Cambridge: Cambridge Scholars Publishing.

Louwerse, M.M. & Van Peer, W. 2006. Waar het over gaat in cijfers. Kwantitatieve benaderuingen in tekst- en literatuurwetenschap. (What it is about in numbers: Quantitative approaches in text- and literary studies). *Tijdschrift voor Nederlandse Taal- en Letterkunde* 122: 21–35.

Louwerse, M.M. & Van Peer, W. In press. How cognitive is cognitive poetics? The interaction between symbolic and embodied cognition. In *Cognitive Poetics*, G. Brone & J. Vandaele (eds). Berlin: Mouton de Gruyter.

Louwerse, M.M. 2004. Semantic variation in idiolect and sociolect: Corpus linguistic evidence from literary texts. *Computers and the Humanities* 38: 207–221.

Manning, C.D. & Schutze, H. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge MA: The MIT Press.

Rowe, J.A. 1988. *Equivocal Endings in Classic American Novels: The Scarlet Letter; Adventures of Huckleberry Finn; The Ambassadors; The Great Gatsby.* Cambridge: CUP.

Sinclair, J. 1966. How to take a poem to pieces. In *Essays on Style and Language*, R. Fowler (ed.), 68–81. London: Routledge and Kegan Paul.

Sinclair, J. 2004. *Trust the Text*. London: Routledge.

Stubbs, M. 2005. Conrad in the computer: Examples of quantitative stylistic methods. *Language and Literature* 14(1): 5–24.

U.S. Census Bureau. 1994. 1990 Census Name Files. Retrieved November 30, 2007, from http://www.census.gov/genealogy/names/names_files.html

Van Peer, W. 1986. Pulp and purpose. Stylistic analysis as an aid to a theory of texts. Linguistics and the study of literature. In *Linguistic Contributions to the Study of Literature*, T. D'Haen (ed.), 268–286. Amsterdam: Rodopi.

Van Peer, W. 1989. Quantitative studies of style: A critique and an outlook. *Computers and the Humanities* 23: 301–307.

Zane, J.P. (ed.). 2007. *The Top Ten: Writers Pick Their Favorite Books*. New York NY: W.W. Norton.