CHAPTER FIVE

UNIGRAMS, BIGRAMS AND LSA: CORPUS LINGUISTICS EXPLORATIONS OF GENRES IN SHAKESPEARE'S PLAYS

MAX LOUWERSE, GWYNETH LEWIS, JIE WU

Abstract

Corpus and computational linguistics could further strengthen the thriving field of empirical studies of literature. This chapter discusses some straightforward corpus linguistic techniques: unigrams, bigrams and latent semantic analysis. The three techniques are then applied to Shakespeare's plays in order to determine how well they can categorize them in genres. In the n-gram analyses frequencies of shared words across the plays are entered in a Multi-Dimensional Scaling (MDS) analysis, in LSA the similarity values between the plays are entered in MDS. With all three techniques two categories emerged: comedies on the one hand, and tragedies/histories on the other. Moreover, a strong correlation was found between the three fundamentally different techniques.

Introduction

The Empirical Study of Literature (ESL) is an interdisciplinary field that draws in researchers from many different fields including Psychology, Sociology, Linguistics, and several others. Since the 1970s, it has attempted to understand the impact of literature on individuals and society. This field has been alluded to as an 'unromantic endeavour' (Cupchik 2007) that demands researchers to consider theories and techniques of a diverse nature in approaching abstract and often challenging problems. Indeed, the ESL endeavour is often less dreamy than the non-empirical approaches one can find in literary studies. And despite the daunting challenges that go hand in hand with understanding the literary phenomena

present in ESL research, the field is nonetheless thriving. It has been growing and branching out into new directions, as is evident in the increase in ESL journal submissions, chapters and volumes dedicated to ESL advancement (e.g. Bortolussi and Dixon 2003; Green, Strange and Brock 2002; Louwerse and van Peer 2002; van Peer, Hakemulder and Zyngier 2007), as well as conferences dedicated solely to ESL research contributions (e.g. IGEL, IAEA). The ratio of publications found in databases dedicated to topics in the field of literary studies between the more empirical PsycINFO and the comparatively less empirical Modern Language Association is steadily increasing, twice as much from the period between 1980-1985 and 1990-1995 and tripled between 1980-1985 and 2000-2005. The year 2006 marked not only the 20th anniversary of The Association for Empirical Studies of Literature (IGEL), but also boasted the highest number ever of contributions from studies in literature, linguistics, psychology, sociology, and computer science.

While the field of ESL has developed greatly since the 1970s, it can be considered a hot-bed of opportunities for the introduction and implementation of research methodology from other areas. At the same time, the term 'empirical studies of literature' has slowly but surely become synonymous with psychological and sociological studies of literature (Schram and Steen 2001). Online and offline experiments measuring reader responses are obviously valuable for the field. Questions of how readers understand literature, how they respond to figurative language, how social processes play a role, all undoubtedly are important in the empirical studies of literature. But the field seems to be losing out on important areas of research that form such an obvious match with the empirical studies of literature, namely that between literary studies and areas of research that investigate texts like corpus linguistics and computational linguistics.

This chapter describes and exemplifies three corpus linguistic techniques available to any ESL researcher: unigram analysis, bigram analysis, and latent semantic analysis. In an exploratory study, plays by Shakespeare are computationally categorized on the basis of their content. The purpose of using the three techniques is to better understand the differences and similarities in their ability to cluster text content, as well as to demonstrate how simple corpus linguistic techniques can be applied to the field of ESL.

Unigrams

One of the simplest and most common methods used in corpus linguistics is the unigram analysis. This technique reduces individual words (unigrams) to their lemmata (e.g. *went* becomes *go*) and compares them to dictionary entries. On the basis of the frequency of these unigrams across texts, it is possible to categorize many different kinds of text (Martindale 1975; Martindale and West 2002), poetry by psychopathological versus non-psychopathological authors was categorized by dividing 3000 words into 36 categories of primordial content. The more primordial content, the more drive- and sensation oriented the text would be. Poetry by psychopathological authors was higher in primordial content. Unigrams can also categorize younger vs. older children's expressive narratives on the basis of primordial content (West, Martindale and Sutton-Smith 1985) as well as the speech of schizophrenic versus non-schizophrenic patients, fantasy stories vs. non-creative stories (Martindale and Daily 1996), and varying themes in narratives (Martindale and West 2002).

Unigrams have even shown to be successful to predict the outbreak of war as demonstrated by Hogenraad (2003) who categorized speeches given by George W. Bush, Tony Blair, and Saddam Hussein throughout the conflict prior to the Iraq war. The trends in words related to affiliation (decreasing) and power (increasing) over time predicted the beginning of the war; (see also the contribution by Hogenraad in this volume).

Along the same lines of research, in a previous study (Louwerse 2004) we attempted to categorize eight Modernist and eight Realist texts using a unigram analysis. Using words from semantic fields that form the Modernist code, such as Observation, Consciousness and Detachment, we tested Fokkema and Ibsch's hypothesis (Fokkema and Ibsch 1988) that the Modernist code is more prevalent in Modernist texts. Results showed differences between authors (idiolects) but not between different codes. That is, frequency of semantic fields did not significantly differ between Modernist texts by Joyce and Woolf on the one hand and Realist texts by Dickens and Eliot on the other.

Unigram analysis is a simple and common approach towards text categorization; however, it also has some drawbacks. Text comparison necessitates the difficult and time-consuming task of building dictionaries. The design of new dictionaries as in the studies discussed so far necessitates answering difficult questions such as what words belong to a certain theme and whether or not the chosen words are unique to that theme.

Another option is to consider all words across texts and to not use dictionaries at all. Such a method would consider all unigrams across texts and look at their differences in frequency. Obviously, it is thereby important to place all texts on equal footing by only selecting those words that can be found across all texts, to avoid a situation whereby a literary work is distinguished solely on the basis of a small set of peculiar words it may use. This technique of comparing frequencies of shared words is particularly useful in those approaches where researchers are mining the data to detect patterns, rather than search for predefined patterns in the data. But a problem using unigram analyses is the sheer number of possible comparisons, since many words are shared across texts.

Another disadvantage of unigrams is that disambiguation of homonyms (i.e. *foot* as a body part vs. *foot* as a measurement) is impossible, so is differentiation between homonyms belonging to different syntactic categories (i.e. John will *ditch* you vs. John fell in a *ditch*).

In sum, unigram analyses have proven to be very useful, but there are a number of drawbacks that should be considered when deciding on this type of analysis.

Bigrams

The drawbacks pointed out in unigram analysis are not unique to corpus linguistics. Speech recognition systems, for instance, share similar problems. Word recognition would be a time consuming process if based on the method of matching spoken words with ones in dictionaries. For every word spoken, the system would have to search through a dictionary to match the first letter, then the second, and so on. This would take a long time because of the endless number of shared first few letters between words. A more efficient method is to use probabilities in predicting a word on the basis of a previous word. Such analysis allows prediction because it is based on frequencies of multiple word combinations instead of one word. This is one of the benefits of using n-gram analysis. Because of size constraints, corpus linguistics generally uses bigrams, two-word combinations.

In previous research (Crossley and Louwerse 2007), we used a bigram analysis to categorize nine spoken and two written corpora. The spoken corpora consisted of the London Lund Corpus (broadcast speeches, face to face conversations, telephone conversations, interviews, spontaneous speech, and prepared speeches), the spatial coordination Map Task Corpus, the temporal coordination TRAINS Corpus, the natural face-to-face conversations in the Santa Barbara Corpus, and the telephone

conversations of the Switchboard Corpus. These corpora were augmented by two written corpora: the Brown Corpus and the Lancaster Oslo Bergen Corpus. Frequencies of the bigrams in each of these 11 corpora were entered in the factor analysis to determine underlying dimensions. Four dimensions emerged:(1) Scripted vs. Unscripted Discourse, separating natural dialogues from monologues, written texts, and tasked-based dialogues; (2) Deliberate vs. Unplanned Discourse, separating written and memorized texts from all other spoken texts; (3) Spatial vs. non-Spatial Discourse, separating the Map Task corpus from all other non-spatial discourse; and (4) Directional vs. non-directional Discourse dominated by the directional and temporal discourse of the TRAINS corpus. What this study shows is that bigram analyses are able to categorize corpora solely on the basis of frequencies of word combinations. The numerous applications to literary works, one explored in this paper, are obvious.

One strength of bigram analyses over unigram analyses is that bigram analyses are not only based on semantic differences. They also reveal latent, syntactic and discourse features. Crossley and Louwerse's (2007) bigram analysis was ostensibly based on lexical collocations, but their bigrams also revealed more than just pure semantics. For instance, the bigram analysis showed that natural, spoken dialogues show a preference for using hedges, a high degree of coordinating conjunctions, especially when combined with first and second person pronouns, the expression of opinions, and a lack of prepositional phrases.

In addition to genre classification, bigrams have also been proven useful in detecting speech acts in dialogue, as shown by (Louwerse and Crossley 2006). Applying an n-gram algorithm to transcribed utterances from the Map Task Corpus (Anderson et al. 1991) resulted in the emergence of content features unique to the Map Task scenario. This study demonstrated that out-of-context textual dialogue acts classification, which can be performed traditionally using state vector machine algorithms, is also feasible using bigram analysis. Furthermore, the performance of such analysis is comparable to that of humans.

While bigrams provide more information than unigrams, the argument could be made that trigrams (three-word combinations) are even more powerful. However, the larger the n-gram, the more sparsity becomes a problem: word frequencies are rarer when they include a greater number of words. For example, the content words *be or not* in *to be or not to be* as quoted in Shakespeare's play *Hamlet* is probably unique to Hamlet (or texts that discuss Hamlet), whereas bigrams like *to be*, *or not*, *to be*, *be or*, *not to* are not. Whereas trigrams present the problem of sparsity, unigrams might yield too much meaningless information (differentiating texts by

frequency of words like *to*, *be*, *or*, *not*). Bigrams, with their word frequencies of two, are an ideal compromise between the two.

The drawback of using bigrams, or n-grams for that matter, is that they rely too much on using the exact words found in the text. Bigrams consider synonyms (e.g. *sad* and *unhappy*) as completely unique to one another. This could be a problem if a number of texts all shared similar themes yet used different but very similar words. A solution to this problem can be found in techniques that focus on latent semantic similarities.

Latent Semantic Analysis

The advantage of using Latent Semantic Analysis (LSA) over unigrams and bigrams is its ability to capture word meanings the relations between synonyms. Word meaning is captured by mapping words into a high dimensional semantic space. The steps include first associating stimuli (words) and their context (in documents), then pairing the associated stimuli on the basis of their contiguity or co-occurrence, and then transforming them using Singular Value Decomposition (SVD) into a smaller number of dimensions (typically 300). This yields a representation of the words, minus the noise. For example, consider:

1) The girl read a book at school.
2) The boy read a book at school.
3) The teacher taught a lesson to the student.

Girl and *boy* are semantically associated based on first-level co-occurrence because they are presented in the same context (they share *a*, *at*, *book*, *school*, *the*). LSA goes farther than relating *girl* and *boy* based only on context. Even though some words may never ever appear in the same document, LSA is still able to map their relations through their semantic neighbours. For instance, the lexical items in sentence 3 (*teacher*, *taught*, *lesson*, *student*) are absent from sentence 1 (*girl*, *read*, *book*, *school*) and sentence 2 (*boy*, *read*, *book*, *school*) but because *student* in sentence 3 shares semantic content with *school* from both sentences 1 and 3, all sentences are therefore semantically related. Note that n-gram techniques (unigram and bigram analyses) are unable to capture any similarity between such sentences.

Various studies have applied the method of statistical knowledge representation. This method has served as an automated essay grader by comparing student essays with ideal essays (Landauer, Foltz and Laham

1998) and performs at a level comparable to students on the TOEFL (Test of English as a foreign language) (Landauer and Dumais 1997). LSA has also been incorporated in software. Coh-Metrix, a web-based tool capable of analyzing text on over 50 types of cohesion relations and over 200 measures of language, text, and readability (Graesser, McNamara, Louwerse, and Cai, 2004; Louwerse, MCarthy, Mcnamara and Graesser 2004; see also Graesser's contribution in the present volume). LSA has also been implemented in intelligent tutoring systems such as AutoTutor, which uses LSA to determine whether a student's answer is more closely related to an ideal, good, or bad answer (Graesser, Lu et al. 2004), and also in iSTART as a basis for appropriate feedback to self-explanations of students (McNamara, Levinstein, and Boonthu 2004).

Recently, Louwerse and van Peer (in press) demonstrated LSA's ability to capture many aspects of cognitive poetics. We took examples from Stockwell (2002) in selecting four topics (figure and ground, prototypes, cognitive deixis, conceptual metaphor) and illustrated how LSA analyses can shed light on the processes of meaning construction.

Although LSA has the advantage of mapping word meanings on many different levels, it has the disadvantage of being less sensitive than unigrams and bigrams to exact words. It also is unable to reveal syntactic and discourse features in text. Bigrams and unigrams, however, reveal such features. Due to these differences, it might be expected that LSA will group text differently than n-grams. The question is, by how much?

Other techniques

Techniques other than n-grams and LSA can be used to analyze literary texts. The appropriate method to apply should be dictated by features of the text in question. For example, Biber (1988) looked at syntactic information instead of the lexical entries that n-grams and LSA consider. He performed a factor analysis on normalized parts-of-speech (special verbs and linguistic constructions) frequencies. This resulted in six factors that showed relations among texts like involved versus informational production, narrative versus non-narrative concerns, explicit versus situation-dependent reference, overt expression of persuasion, abstract vs. non-abstract information and on-line informational elaboration. It was found that romantic fiction, mystery fiction, and science fiction were categorized under narrative, while academic prose and official documents were categorized as belonging to a non-narrative dimension. As of late, Biber's multi-feature, multi-dimensional approach has become a standard

in corpus linguistics (McEnery and Wilson 2001) and has led to extensions on the method (Biber, Conrad, and Reppen 1994; Biber and Conrad 2001).

Linguistic features of a great variety can be identified at word level (e.g. morpho-semantics, syntactic category, frequency) in text. As Biber's 1988 study demonstrates, such identifiable linguistic features serve as powerful determiners of similarities and differences between registers. While such features have yielded persuasive results, we still do not know how effective these are in capturing the nature of a register, because we do not know if it effectively captures the nature of a text. Certainly, linguistic features may betray several register characteristics at a word level, but what does it tell us about the meaning of the text as a whole? Without understanding the structure of words in sentences and sentences in paragraphs etc, we do not have an understanding of the text. How can we then understand the processes behind the text? Readers comprehend text by actively constructing a coherent mental representation of the information presented in it. This coherent representation is built from textual indications that form a text's cohesion (Louwerse and Graesser 2005). This cohesion can not be understood solely at the word level. It instead necessitates understanding at the textual levels of inter-clause, inter-sentence, and inter-paragraph structure.

Text cohesion can be determined by means of the previously mentioned web-based tool Coh-Metrix (McNamara, Louwerse, and Graesser 2002; Graesser, McNamara, et al. 2004). This tool analyzes texts on over 230 types of cohesion relations and measures of language, text and readability through modules that use lexicons, parts-of-speech classifiers, syntactic parsers, templates, corpora, latent semantic analysis, and other widely used components in computational linguistics. These other components include WordNet (Miller et al. 1990), which determines underlying lexical concepts; the MRC database (Coltheart 1981) for gathering psycholinguistic information; Latent Semantic Analysis (Landauer and Dumais 1997) for semantic similarities between words, sentences, and paragraphs; the ApplePie parser (Sekine and Grishman 1995); and the parts-of-speech tagger (Brill 1995) for various syntactic categories. For further information about Coh-Metrix's measures, see Graesser, McNamara et al. (2004).

Coh-Metrix is readily available to researchers at http://coh-metrix.memphis. edu. Biber's software is not yet available to the public (see Biber 1988 for a detailed description of the software's algorithm). Louwerse et al. (2004) investigated register variations using Biber's approach in combination with Coh-Metrix.

The focus of this paper is on unigram, bigram, and LSA techniques, as they are most accessible (unigrams and bigrams can be computed using software available at http://www.madresearchlab.org and LSA is available at http://lsa.colorado.edu). Secondly, these methods all focus on one measure (lexical collocations) rather than a multitude of measures. Finally, these measures have been used in our previous work (Crossley and Louwerse 2007; Louwerse and Crossley 2006; Louwerse et al. 2004; Louwerse and Van Peer in press). With this in mind, the three methods in question are used to investigate clustering Shakespearean texts.

Categorizing Shakespeare's Plays

We investigated classifications of Shakespearean texts using the three methods described earlier: unigram, bigram and LSA analysis. For this purpose, we used all of the 37 of Shakespeare's plays. These texts were originally (in the Folio of 1623) categorized under comedies, histories and tragedies.

Table 1: *Canonical categorization Shakespeare plays*

Comedy	History	Tragedy
All's Well That Ends Well	Cymbeline	Antony and
As You Like It	Henry IV, part I	Cleopatra
Comedy of Errors	Henry IV, part II	Coriolanus
Love's Labour's Lost	Henry V	Hamlet
Measure for Measure	Henry VI, part I	Julius Caesar
Merry Wives of Windsor	Henry VI, part II	King Lear
Merchant of Venice	Henry VI, part III	Macbeth
Midsummer Night's Dream	Henry VIII	Othello
Much Ado About Nothing	King John	Romeo and Juliet
Pericles	Richard II	Timon of Athens
Taming of the Shrew	Richard III	Titus Andronicus
Tempest		
Troilus and Cressida		
Twelfth Night		
Two Gentlemen of Verona		
Winter's Tale		

It is important to note here that genre clustering in Shakespeare's plays is typically based on plot. Holzknecht (1950) shows that tragedies are tragedies because they start with happiness and prosperity and end with sadness and death; comedies start with all things wrong, and end with all happiness being restored. Historical plays, finally, were an accidental form, closely akin to tragedy (idem, p. 251). Plot, however, is not captured by either n-grams or LSA. Because none of the categorizations are based on Shakespeare's language use (cf. De Grazia 2001), there is no evidence that these techniques can capture genres in Shakespeare's plays.

An additional problem concerns defining the ultimate 'correct' categorization. Comedies like *All's Well That Ends Well*, *Measure for Measure* and *Troilus and Cressida* have been called 'problem plays', or 'black comedies', because they have both happy and frivolous as well as dark and violent elements (Boas 1896; Snyder 2001). Other plays, including *Hamlet*, *The Winter's Tale*, *Timon of Athens*, and *The Merchant of Venice* have also been considered problem plays, because they do not fit the plot of tragedies well. Other plays, like *Cymbeline*, once called tragedies, are later categorized as comedies, but have also been categorized as romance. Additional plays considered romances are *The Tempest*, *The Winter's Tale*, *Pericles* and *Two Noble Kinsmen* (Campbell 1966).

In sum, genre classification of Shakespeare's plays is difficult for the human expert despite the vast contextual and historical resources available to base classification on. Corpus linguistic techniques, which rely on only lexical items to distinguish genres, are therefore expected to pose a degree of challenge.

Unigrams and Shakespeare's Plays

Word frequency lists were computed for all 37 texts whereby only those words that were found across all texts were included. Frequencies were normalized for each text by computing percentiles. This resulted in frequency lists of a total of 174 words. Not surprisingly, this list contained high-frequency items that can be found in typical word frequency lists like Kucera and Francis (1967) and CELEX (Baayen, Piepenbrock, and van Rijn 1993), including words like *a*, *and*, *did*, *in*, *of*, *that*, *the*, *to*. Because we are primarily interested in underlying dimensions of the relationships between the texts, in this and subsequent analyses, these normalized frequencies were then supplied to an ALSCAL algorithm to derive a Multidimensional Scaling (MDS) representation of the stimuli (Kruskal and Wish 1978). That is, from the normalized frequencies we computed a

matrix of Euclidean distances. This matrix was compared with arbitrary coordinates in an n-dimensional space. The coordinates were iteratively adjusted such that Kruskal's stress is minimized and the degree of correspondence maximized: High stress and lower correspondence indicates a poor fitting of the data, whereas a low stress and high correspondence indicates a good fitting.

The fitting of the data in two dimensions was satisfactory (Kruskal's stress 1= .28; R^2= .74). For the purpose of this chapter we will focus on one of the two dimensions that emerged from the data and separated Shakespeare's comedies from his tragedies and histories. Coordinates for each text are given in Table 2.

Table 2: *Positioning of Shakespeare text on the dimension comedy versus history/tragedy (unigrams).*

Title	Genre	Coordinates
Merry Wives of Windsor	comedy	-2.445
Much Ado About Nothing	comedy	-2.017
The Two Gentlemen of Verona	comedy	-1.957
The Taming of The Shrew	comedy	-1.913
Twelfth Night	comedy	-1.818
As You Like It	comedy	-1.473
Two Noble Kinsmen	comedy	-1.462
Tragedy of Macbeth	tragedy	-1.391
Measure For Measure	comedy	-1.370
All's Well That Ends Well	comedy	-1.365
The Winter's Tale	comedy	-0.999
The Merchant of Venice	comedy	-0.982
Cymbeline	comedy	-0.485
Hamlet	tragedy	-0.351
Julius Caesar	tragedy	-0.317
Antony And Cleopatra	tragedy	-0.298
The Comedy of Errors	comedy	-0.245
The Tempest	comedy	-0.227
Othello	tragedy	-0.155
A Midsummer Night's Dream	comedy	-0.036
Troilus And Cressida	history	0.064
Titus Andronicus	tragedy	0.070
Henry IV, part II	history	0.096

King Henry V	history	0.109
King Henry VIII	history	0.170
Love's Labour's Lost	comedy	0.353
Coriolanus	tragedy	0.465
Henry VI, part I	history	0.506
King Richard II	history	0.745
Romeo and Juliet	tragedy	1.399
King Lear	tragedy	1.408
Henry VI, part II	history	1.709
King John	history	1.751
King Henry VI, part III	history	1.754
King Richard III	history	1.769
Henry IV, part I	history	1.834
Timon of Athens	tragedy	1.844

As the table shows, the classification of texts according to genre on the basis of frequencies of shared words alone is quite accurate. The two exceptions are *Macbeth* which is erroneously classified as a comedy, and *Love's Labour's Lost* which is erroneously classified as a history/tragedy.

The distinction between comedies on the one hand and tragedies and histories on the other may seem puzzling at first. But when one considers genre classification in Shakespeare's plays, the distinction is not surprising. Snyder (2001), for instance, refers to Francis Meres who in 1598 initially listed *Richard II*, *Richard III*, *King John* and *Henry IV* as tragedies, but later classified them as histories.

This classification allows us to determine which plays are considered to be prototypical for the different genres. According to the analysis a prototypical comedy is *The Merry Wives of Windsor*, prototypical histories and tragedies *Henry IV, part I*, and *Timon of Athens*.

What are the unigrams that best distinguish Comedies on the one hand and the Tragedies/Histories on the other? Table 3 presents a list of some of the words indicative for either category. For instance, words like *of, the, and, in, his, to, all, thou, that, there are* most frequent in comedies and least frequent in tragedies/histories.

Table 3: *Unigrams most indicative of comedy versus tragedy/history*

Comedy	Tragedy/History
of, the, and, in, his, to, all, thou, that, their, did, my, thy, o, which, from, was, our, by, thee	with, no, good, if, as, not, have, love, your, be, it, she, me, for, will, her, is, a, you, I

As we argued before, in unigram analyses it is hard to explain why certain words occur relatively more often in one text than another. Tragedies and histories can be marked by a higher frequency of pronouns (*I*, *you*, *she*, *it*) as well as lexical items (*no*, *good*, *love*). Grammatical items can be found more frequently in comedies (all indicators of this genre are grammatical items).

What we can conclude from this analysis is that a simple unigram analysis, whereby only the frequencies of words are considered, allows for quite an accurate classification of Shakespeare's plays into two categories that have been considered in the literature: comedies on the one hand and tragedies/histories on the other[1].

Bigrams and Shakespeare's plays

The drawback of MDS is that it remains exploratory and the current clustering may consequently be coincidental. We therefore extended the unigram analysis to a bigram analysis, otherwise following the same procedure.

The bigram analysis replicated the unigram analysis except that lexical items consisted of two-word combinations. A total of 77 bigrams were

[1] At the end of this analysis it may be worthwhile to spend a few words on the MDS method we have used. Because our goal is to detect meaningful underlying dimensions that allow us to explain similarities and dissimilarities between words, and because of the exploratory character of this study, we believe that an exploratory statistical measure like MDS is warranted. It may in fact be worth motivating the choice for MDS over Factor Analysis. Guttman (1977) showed the similarities between factor analysis and MDS. Indeed, when unigram results for a factor analysis (Varimax rotation) are compared with those for the MDS, correlations follow identical patterns, with correlations at $r = .79$, $p < .001$. The advantage of MDS over Factor Analysis lies in the presentation and interpretation and is therefore chosen here.

found that were the same across all 37 texts. Their frequencies were again normalized by text size. Some of the frequent bigrams were *in his*, *of his*, *in the*, *of the*, *by the*, *to your*, *and to*, *with his* and *I have*. Normalized frequencies for all texts were again entered in an MDS. The fitting of the data in two dimensions was acceptable (Kruskal's stress 1= .29; R^2= .65).

Interestingly, as Table 4 shows, comedies were again quite reliably distinguished from tragedies/histories. Exceptions were *Antony and Cleopatra*, which was classified as a comedy and (again) *Love's Labour's Lost*, identified as a tragedy/history. The similarities in categorizations between unigram and bigram analyses are obvious (r = .6, p < .001, N = 37).

Table 4: *Positioning of Shakespearean text on the dimension comedy versus history/tragedy (bigrams).*

Title	Genre	Coordinates
Twelfth Night	comedy	-2.323
Antony And Cleopatra	tragedy	-1.941
The Taming of The Shrew	comedy	-1.734
The Merchant of Venice	comedy	-1.683
Cymbeline	comedy	-1.599
The Winter's Tale	comedy	-1.588
Much Ado About Nothing	comedy	-1.579
All's Well That Ends Well	comedy	-1.563
The Two Gentlemen of Verona	comedy	-1.459
The Merry Wives of Windsor	comedy	-1.327
As You Like It	comedy	-1.096
The Two Noble Kinsmen	comedy	-1.089
Julius Caesar	tragedy	-0.678
King Henry V	history	-0.555
The Comedy of Errors	comedy	-0.517
Hamlet	tragedy	-0.501
Measure For Measure	comedy	-0.313
The Tempest	comedy	-0.267
Macbeth	tragedy	-0.265
A Midsummer Night's Dream	comedy	-0.150
King Henry VIII	history	-0.054
Henry VI, part I	history	0.142
King Lear	tragedy	0.197

Henry IV, part II	history	0.282
Love's Labour's Lost	comedy	0.677
Titus Andronicus	tragedy	0.718
Romeo and Juliet	tragedy	0.723
Othello	tragedy	0.733
Coriolanus	tragedy	0.828
King Richard III	history	1.037
King John	history	1.082
King Henry VI, Part II	history	1.135
Henry IV, part I	history	1.210
King Richard II	history	1.351
Troilus and Cressida	history	1.431
King Henry VI, part III	history	1.813
Timon of Athens	tragedy	2.219

The most indicative bigrams for the two categories are presented in Table 5. As in the unigrams of the previous analysis, tragedies and histories seem to be more self-centered than comedies. Bigrams like *I had*, *that I*, *I have*, *I am* mark tragedies/histories, bigrams like *of his*, *for his*, *in his*, *with his* mark comedies. One explanation for these findings can come from social psychology. Stirman and Pennebaker (2001) found that suicidal poets use language that was more concerned with the self. The bigram analysis here suggests that we cry about ourselves, we laugh about others.

Table 5: *Bigrams most indicative of comedy versus tragedy/history*

Comedy	Tragedy/History
to the, thou art, with the, of his, from the, my heart, for his, and all, in his, of a, me and, with his, to my, to make, to me, and in, but to, in a , to your, would have	when he, that is , be so, you and, I had, by the, and the, that I , such a, will not, in the , in my, and so, to be, of the, you are, of my, it is, I have, I am

N-gram analyses rely on exact words in the text. The argument can be made that the extreme differences in frequencies of certain words that happen to occur in all texts have supported the categorization into the two genre groups. This argument is not a strong one for a number of reasons: only those words that appeared in all texts were selected and those words were typically high-frequency words. Moreover, the MDS analysis is not sensitive to a small set of differences but scales on the basis of the normalized frequencies of all words. Finally, our previous work, most notably Crossley and Louwerse (2007) and Louwerse and Crossley (2006) have provided evidence for the use of bigrams in categorizations. However, to further support the evidence that corpus linguistic measures can categorize Shakespeare's plays we used the third technique discussed in this chapter, one that is not sensitive to particular words, Latent Semantic Analysis.

Latent Semantic Analysis and Shakespeare's plays

For the unigram analysis, all texts were computed for frequencies of the unigrams shared across all texts. A Multidimensional Scaling (MDS) algorithm was then applied to the frequencies. In the bigram analysis the method was identical to the unigram analysis, except that frequencies of pairs were used. In the LSA analysis, we used a 37 x 37 cosine value matrix entered into MDS whereby each cosine value represented a semantic similarity between plays.

For the LSA analyses a 'knowledge base' is needed in the form of an LSA space. We used the Touchstone Applied Science Associates (TASA) Corpus. 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 (Kintsch 1998; Landauer and Dumais 1997). The immediate argument against this corpus is that it is modern discourse, with expository and narrative texts - not particularly Shakespeare's language. The strength of LSA, however, is its low sensitivity to the semantic space, meaning it can still function adequately despite the mismatch. Besides, if the LSA space were to be a problem, a poor fitting of the data and a classification very different from the previous ones would be predicted.

The matrix of 37 x 37 LSA cosine values was supplied to an ALSCAL algorithm for an MDS representation. The fitting of the data in two

dimensions was good (Kruskal's stress 1= .15; R^2= .96). The dimension of interest here is the one we discussed earlier. As with the unigram and bigram analyses, LSA plotted histories/tragedies versus comedies in separate categories. That categorization is not perfect (*Romeo and Juliet* and *Antony and Cleopatra* are considered comedies; *The Tempest* and *Love's Labour's Lost* are considered history/tragedy), but overall the distinction is accurate, as shown in Table 6.

Love's Labour's Lost was categorized all three times (unigram, bigram and LSA) as a history/tragedy. Before we simply label this as a miscategorization, it is interesting to note that the play has many comical events, but does end with a much darker tone.

Table 6: *Positioning of Shakespeare text on the dimension comedy versus history/tragedy (LSA).*

Title	Genre	Coordinate
King Richard II	history	-1.661
King Henry VI, part III	history	-1.651
King Henry VI, Part II	history	-1.478
King Richard III	history	-1.362
The Henry IV, part I	history	-1.317
Timon of Athens	tragedy	-1.053
King John	history	-1.031
Henry VI, part I	history	-0.959
Othello	tragedy	-0.724
Coriolanus	tragedy	-0.597
King Henry VIII	history	-0.560
The Tempest	comedy	-0.482
Henry IV, part II	history	-0.398
Love's Labour's Lost	comedy	-0.397
King Henry V	history	-0.368
Titus Andronicus	tragedy	-0.314
Macbeth	tragedy	-0.186
Julius Caesar	tragedy	-0.148
Cymbeline	comedy	-0.126
Troilus and Cressida	history	0.023
All's Well That Ends Well	comedy	0.067
The Winter's Tale	comedy	0.163

King Lear	tragedy	0.188
Hamlet	tragedy	0.388
Two Noble Kinsmen	comedy	0.391
The Merchant of Venice	comedy	0.400
A Midsummer Night's Dream	comedy	0.461
The Comedy of Errors	comedy	0.536
Romeo And Juliet	tragedy	0.692
Antony and Cleopatra	tragedy	0.865
As You Like It	comedy	0.887
Twelfth Night	comedy	1.003
Measure for Measure	comedy	1.146
Much Ado About Nothing	comedy	1.420
The Taming of The Shrew	comedy	1.479
The Two Gentlemen of Verona	comedy	1.628
The Merry Wives of Windsor	comedy	2.003

LSA analyses do not provide insight in how corpora are clustered, simply because they use higher-order relationships. It is noteworthy that fundamentally different techniques, bigrams and LSA, yield very similar results. The bigram analysis typically includes high-frequency grammatical items (non-lexical items) like *and, I, have, in*. The strength of LSA, on the other hand, lies in not-so-high frequent lexical items. Nevertheless, LSA results correlate with unigram (r = .86, p < .001, N = 37) as well as with bigram results (r = .59, p < .001, N = 37).

Conclusion

The current chapter has provided a brief description of three techniques that have been used frequently in corpus linguistics. We have selected rather straightforward techniques, available to any Humanistic researcher, and have shown how they can be used, for instance, in categorizing Shakespeare's plays into genres.

The field of empirical studies of literature is thriving, with more observable, reliable and valid results being reported, the present volume being an excellent example. At the same time, ESL has moved more and more towards cognitive and social psychology and sociology. Our recommendation for a new beginning for the study of literature is therefore that ESL keeps considering the wide interdisciplinary spectrum of areas

and approaches. This should at least include obvious ones like computational linguistics and corpus linguistics: two bigrams, sharing the same unigram, with a high LSA match.

Works Cited

Anderson, A., M. Bader, E. Bard, E. Boyle, G. M. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. S. Thompson, and R. Weinert. 1991. The HCRC Map Task Corpus. *Language and Speech* 34: 351-366.

Baayen, H., R. Piepenbrock, and H. van Rijn. 1993. The CELEX Lexical Database. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Biber, D. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D., and S. Conrad. 2001. Quantitative corpus-based research: Much more than bean counting. *TESOL Quarterly* 35: 331-6.

Biber, D., S. Conrad, and R. Reppen. 1994. Corpus-based approaches to issues in applied linguistics. *Applied Linguistics* 15: 169-189.

Boas, F. S. 1896. *Shakespeare and his predecessors.* New York: Charles Scribner's Sons.

Bortolussi, M., and P. Dixon. 2003. *Psychonarratology: Foundations for the empirical study of literary response.* Cambridge: Cambridge University Press.

Brill, E. 1995. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics* 21: 543-566.

Campbell, O. J. 1966. *The Reader's encyclopedia of Shakespeare*. New York: Crowell.

Coltheart, M. 1981. The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology* 33: 497-505.

Crossley, S. A., and M. Louwerse. 2007. Multi-dimensional register classification using bigrams. *International Journal of Corpus Linguistics* 4: 453-478.

Cupchik, G. 2007. Finding an empirical link between psychology and the humanities. *International Society for the Empirical Study of Literature Newsletter* 2007 no. 17. http://www.arts.ualberta.ca/igel/Newsletter17 Jan 29.

De Grazia, M. 2001. Shakespeare and the craft of language. In *The Cambridge Companion to Shakespeare*, edited by M. de Grazia and S. Wells, 49-64. Cambridge: Cambridge University Press.

Fokkema, D. W., and E. Ibsch. 1988. *Modernist conjectures: A mainstream in European literature*. New York: St. Martin's Press.

Graesser, A. C., D. S. McNamara, M. M. Louwerse, and Z. Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, and Computers* 36: 193-202.

Graesser, A. C., S. Lu, G. T. Jackson, H. H. Mitchell, M. Ventura, A. Olney, and M. M. Louwerse. 2004. AutoTutor: A tutor with dialogue in natural language. *Behavioral Research Methods, Instruments, and Computers* 36: 180–93.

Green, M. C., J. J. Strange, and T. C. Brock, eds. 2002. *Narrative impact: Social and cognitive foundations*. Mahwah, NJ: Lawrence Erlbaum Associates. Inc.

Guttman, L. 1977. What is not what in statistics? *The Statistician* 26: 81-107.

Hogenraad, R. 2003. The words that predict the outbreak of wars. *Empirical Studies of the Arts* 21: 5–20.

Holzknecht, K. J. 1950. *The backgrounds of Shakespeare's plays*. New York: American Book.

Kintsch, W. 1998. *Comprehension: A paradigm for cognition*. Cambridge, UK: Cambridge University Press.

Kruskal, B. J., and M. Wish. 1978. *Multidimensional scaling*. Beverley Hills, CA: Sage Publications.

Kucera, H., and W. N. Francis. 1967. *Computational analysis of present-day English*. Providence, RI: Brown University Press.

Landauer, T. K., Foltz, P. W., and D. Laham. 1988. Introduction to latent semantic analysis. *Discourse Processes* 25: 259-284

Landauer, T. K., and S. T. Dumais. 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.

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

Louwerse, M. M., and S. Crossley. 2006. Dialog act classification using n-gram algorithms. In *Proceedings of the* 19*th International Florida Artificial Intelligence Research Society*, 758-763. Menlo Park, CA: AAAI Press.

Louwerse, M. M., & A. C. Graesser. (2005). Coherence in discourse. In *Encyclopedia of linguistics*, edited by P. Strazny,, 216-218. Chicago: Fitzroy Dearborn.

Louwerse, M. M, P. M. McCarthy, D. S. McNamara, and A. C. Graesser. 2004. Variation in language and cohesion across written and spoken

registers. In *Proceedings of the* 26*th Annual Meeting of the Cognitive Science Society*, edited by K. Forbus, D. Gentner, and T. Regier, 843-848. Mahwah, NJ: Erlbaum.

Louwerse, M. M, and W. van Peer, eds. 2002. *Thematics: Interdisciplinary studies*. Amsterdam/Philadelphia: John Benjamins.

Louwerse, M. M., and W. V. Peer. 2006. *Waar het over gaat in cijfers. Kwantitatieve benaderuingen in tekst- en literatuurwetenschap*. [What numbers are about: quantitative approaches in text- and literary studies]. *Tijdschrift voor Nederlandse Taal- en Letterkunde* 122: 21-35.

Louwerse, M. M., and W. V. Peer. in press. How cognitive is cognitive poetics? The interaction between symbolic and embodied cognition. In *Cognitive Poetics*, edited by G. Brone and J. Vandaele. Berlin: De Gruyter.

Martindale, C. 1975. *Romantic progression: The psychology of literary history*. Washington, DC: Hemisphere.

Martindale, C., and A. Daily. 1996. Creativity, primary process cognition, and personality. *Personality and Individual Differences* 20: 409-414.

Martindale, C., and A. West. 2002. Quantitative hermeneutics. Inferring meaning of narratives from trends in their content. In *Thematics: Interdisciplinary Studies*, edited by M. M. Louwerse and W. van Peer, 377-395. Amsterdam: Benjamins.

McEnery, A. M., & A. Wilson. 2001. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

McNamara, D. S., H. Levinstein, and C. Boonthum. 2004. iSTART: Interactive Strategy Training for Active Reading and Thinking. *Behavioral Research Methods, Instruments, and Computers* 36: 222 - 233.

McNamara, D. S., M. M. Louwerse, and A. C. Graesser. 2002. Coh-Metrix: Automated cohesion and coherence scores to predict readability and facilitate comprehension. Unpublished technical report: University of Memphis.

Miller, G., R. Beckwith, R. Fellbaum, and D. Gross. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4): 235-244. Oxford University Press.

Schram, D. H., and G. J. Steen, eds. 2001. *The psychology and sociology of literature*. Amsterdam: John Benjamins.

Sekine, S., and R. Grishman. 1995. A corpus-based probabilistic grammar with only two non-terminals. *Fourth International Workshop on Parsing Technology*, 260-270. Prague: Karlovy Vary

Snyder, S. 2001. The genres of Shakespeare's plays. In *The Cambridge Companion to Shakespeare*, edited by M. de Grazia and S. Wells, 83-97. Cambridge: Cambridge University Press.

Stirman, S. W., and J. W. Pennebaker. 2001. Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic Medicine* 63: 517-522.

Stockwell. P. 2002. *Cognitive poetics: An introduction*. London: Routledge.

van Peer, W., J. Hakemulder, and S. Zyngier. 2007. *Muses and measures: Empirical research methods for the humanities.* Newcastle-upon-Tyne: Cambridge Scholars Publications.

West, A., C. Martindale, and B. Sutton-Smith. 1985. Age trends in the content of children's spontaneous fantasy narratives. *Genetic, Social, and General Psychology Monographs* 111: 391-405.

SHORT BIOS