Stylogenetics: Clustering-based stylistic analysis of literary corpora

Kim Luyckx, Walter Daelemans, Edward Vanhoutte

University of Antwerp
Faculty of Arts
Universiteitsplein 1
B-2610 Antwerp
Belgium
{kim.luyckx, walter.daelemans, edward.vanhoutte}@ua.ac.be

Abstract

Current advances in shallow parsing allow us to use results from this field in stylogenetic research, so that a new methodology for the automatic analysis of literary texts can be developed. The main pillars of this methodology - which is borrowed from topic detection research - are (i) using more complex features than the simple lexical features suggested by traditional approaches, (ii) using authors or groups of authors as a prediction class, and (iii) using clustering methods to indicate the differences and similarities between authors (i.e. stylogenetics). On the basis of the stylistic genome of authors, we try to cluster them into closely related and meaningful groups. We report on experiments with a literary corpus of five million words consisting of representative samples of female and male authors. Combinations of syntactic, token-based and lexical features constitute a profile that characterizes the style of an author. The stylogenetics methodology opens up new perspectives for literary analysis, enabling and necessitating close cooperation between literary scholars and computational linguists.

1. Introduction

Recently, language technology has progressed to a state of the art in which robust and fairly accurate linguistic analysis of lexical, morphological, and syntactic properties of text has become feasible. This enables the systematic study of the variation of these linguistic properties in texts by different authors (author identification) (Baayen et al., 1996; Gamon, 2004), different time periods, different genres or registers (Argamon et al., 2003), different regiolects, and even different genders (Koppel et al., 2003; Kelih et al., 2005).

We see this trend as potentially providing new tools and a new methodology for the analysis of literary texts that has traditionally focused on complex and deep markup (McCarty, 2003) and the statistical assesments of concordances and word-count applications (Raben, 1965; Burrows, 1987; Lancashire, 1993; Bucher-Gillmayr, 1996) for the analysis of rhyme and sound patterns (Wisbey, 1971; Robey, 2000), the investigation of imagery and themes (Corns, 1982; Fortier, 1989; Fortier, 1996; Ide, 1986; Ide, 1989), the structure of dramatic works (Potter, 1981; Potter, 1989; Steele, 1991; Ilsemann, 1995), stylometrics and authorship attribution (Hockey, 2000, 104-123), (Craig, 2004). See (Rommel, 2004) for an overview of computational methods in literary studies. The methodology we propose is borrowed from the text categorization literature (Sebastiani, 2002) where simple lexical features (called a bag of words) are used to characterize a document with some topic class. Statistical and information-theoretic methods are used to select an informative bag of words to distinguish between documents with different topics. Machine Learning methods are then used to learn to assign documents to one of the predefined topics on the basis of examples. We generalize this methodology in three ways:

- i. By extending the simple lexical features with more complex features based on distributional syntactic information about part of speech tags, nominal and verbal constituent patterns, as well as features representing readability aspects (average word and sentence length, type/token ratio etc.). The statistical and information-theoretic methods can then be applied to more complex features than individual words for stylistic analysis.

- ii. By using individual authors or groups of authors as classes to be predicted rather than topics. It can then be investigated which features are predictive for author identity, gender, time period etc. See (Koppel et al., 2003) for work on this approach for gender prediction.

- iii. By using the vectors of complex features, computed on a sufficiently large sample of the work of an author as a signature for the style of that author and using similarity-based clustering methods to develop a stylogenetic analysis of differences and similarities between authors, periods and genders. We define stylogenetics here as an approach to literary analysis that groups authors on the basis of its stylistic genome into family trees or closely related groups from some perspective.

Tree classification as a tool for the study of proximity and distance between texts and authors has recently been explored by few studies which take the whole vocabulary of the texts which are compared into consideration. (Julliard and Luong, 1997; Julliard and Luong, 2001; Spencer et al., 2003; Labbé and Labbé, to appear 2006). Central in these studies, however, are not the complex features as proposed in our methodology, but the lexical and lexicographical standardization of the vocabulary that is the qualitative basis for proximity measurements between pairs of texts.

2. Corpus

In this paper we report on explorative stylogenetic work using a large corpus of literary works. From three online text archives (viz. The Oxford Text Archive, the Electronic Text Center of the University of Virginia and to a minor extent Project Gutenberg) we collected representative samples of 100,000 words of 50 English and American authors, half of them male, half of them female, from 12 time periods between 1525 and 1925 (we worked with 25-year periods). The appendix provides an overview of the authors, genders, and periodization of the samples used (cf. Tables 1, 2).

3. Feature Extraction

Four types of features that have been applied as style markers can be distinguished: token-level features (e.g. word length, readability), syntactic features (e.g. part-of-speech tags, chunks), features based on vocabulary richness (e.g. type-token ratio) and common word frequencies (e.g. of function words) (Stamatatos et al., 2001). While most stylometric studies are based on token-level features, word forms and their frequencies of occurrence, syntactic features have been proposed as more reliable style markers since they are not under the conscious control of the author (Baayen et al., 1996; Diederich et al., 2000; Khmelev and Tweedie, 2001; Kukushkina et al., 2001; Stamatatos et al., 1999). Thanks to improvements in shallow text analysis, we can extract syntactic features to test their relevance in stylogenetic research.

In a first step, we developed an environment which enables the automatic production of profiles of the samples in the Stylogene corpus. A profile consists of a vector of 208 numerical features representing automatically assigned information about the following features:

- **Type-token ratio**: The type-token ratio *V/N*, *V* representing the size of the vocabulary of the sample, and *N* the number of tokens, is a measure indicating the vocabulary richness of an author.

- **Word length**: The distribution of words of different lengths has been used as a feature in authorship attribution studies (Diederich et al., 2000). Words with a length of 15-19, 20-24 and 25+ were combined in separate categories.

- **Readability**: The readability feature is an implementation of the Flesch-Kincaid metric which indicates the readability of a text, using mean word and sentence length.

- **Distribution of parts-of-speech**: Syntax-based features are not under the conscious control of the author and therefore reliable style markers. Somers suggests that

  > A more cultivated intellectual habit of thinking can increase the number of substantives used, while a more dynamic empathy and active attitude can be habitually expressed by means of an increased number of verbs. (Holmes, 1994, 89)

- **Distribution of frequent function words**: Traditional approaches to stylometry research use content words rather than function words, assuming that the latter occur to frequently to be of any relevance for style. Nevertheless, function words (e.g. determiners, conjunctions, prepositions) are not under the conscious control of the author and therefore meaningful for stylogenetic studies (Holmes, 1994, 90-91).

- **Distribution of frequent chunks**: Similarly to parts-of-speech, chunks are also reliable features for stylogenetic research. We automatically extracted frequencies of noun phrase, verb phrase, prepositional phrase, adjectival phrase, adverbial phrase, conjunction, interjection, verb particle, subordinated clause and preposition-noun phrase chunks.

- **NP and VP chunk internal variation**: The internal organisation of NP and VP chunks is subject to variation, which can reveal the subconscious preference of the author.

The resulting profiles can be used in applications like author or gender identification, but also in a stylogenetic analysis for the discovery of stylistic relationships between authors that may not be evident on the basis of a more superficial comparison. As a representation of contemporary non-literary language, we added a profile based on 100,000 words of Wall Street Journal text.

In order to be able to extract these features automatically, we used shallow parsing software developed in our lab (Daelemans and van den Bosch, 2005) to automatically assign parts of speech and constituent structure to the 51 x 100,000 word corpus. The pos tag set and chunk label set used are those of the Penn Treebank project (Marcus et al., 1993).

4. Cluster Analysis and Interpretation

The clustering method used is the one implemented in the *cluster* program of Andreas Stolcke, which is an instance of Euclidean distance based centroid clustering. Initially, all data points are treated as clusters and the most similar clusters are iteratively merged into larger clusters, building up a hierarchical tree.

Figure 1 shows the family tree produced by applying hierarchical clustering with Euclidean distance as similarity metric to the full profiles of each author.

In further exploratory research, we used information-theoretic analysis (i.e. Gain Ratio) of the relevance of each feature in the profile in predicting the gender of the author as a heuristic to select a new profile to cluster for gender-related stylistic family trees. We selected the 43 features that turned out to be the most relevant for characterizing style differences between genders.

Figure 2 shows the family tree after feature selection in which we find five groups of gender clusters.

The tree in Figure 1 shows that the Wall Street Journal (WSJ) profile is clearly separated from the rest of the corpus and that within the latter, Defoe, Hobbes, Mill, Behn, and More are stylistic outliers. The interrelation between genre and period may explain their distance from the rest

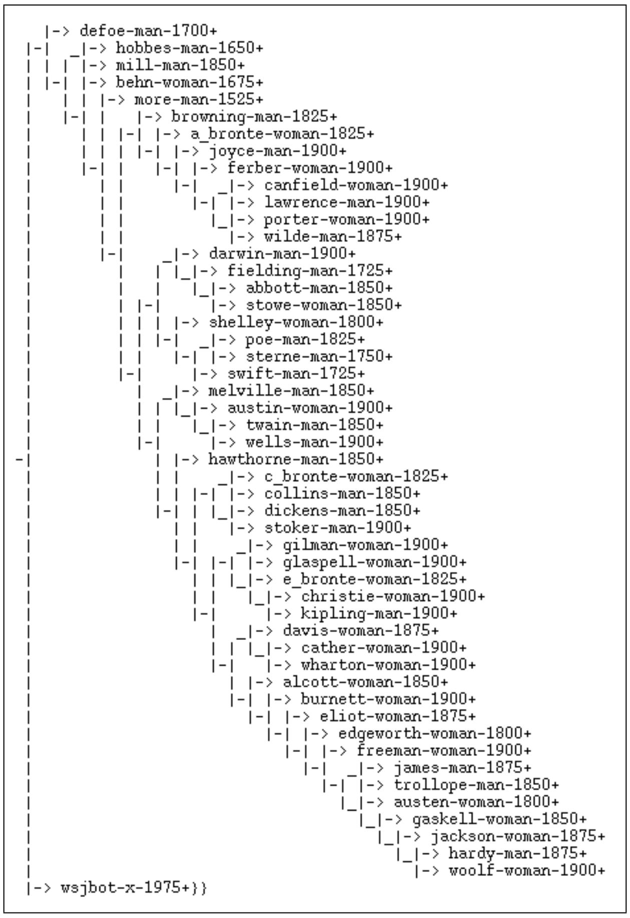

Figure 1: Family tree based on entire feature set

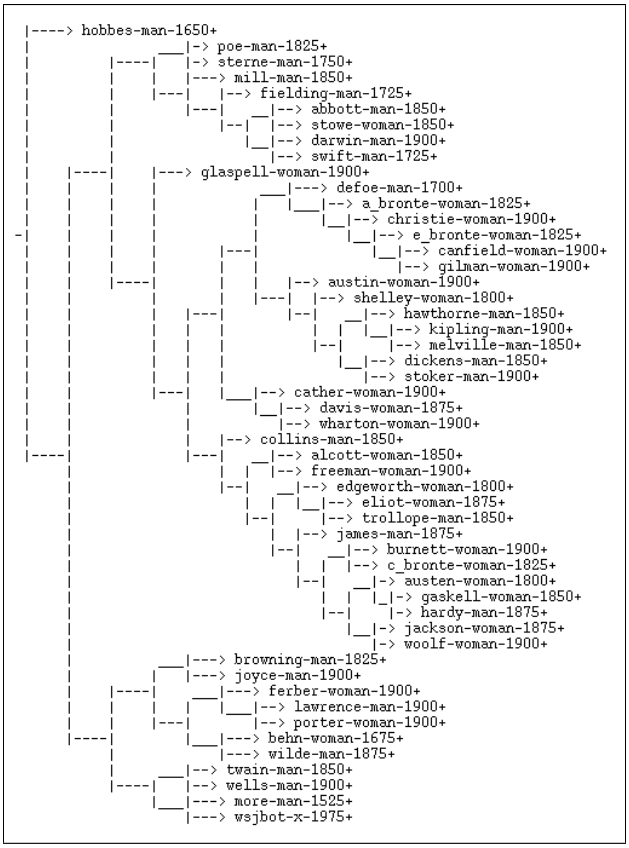

Figure 2: Family tree after feature selection on gender clustering

of the stylogene corpus. Hobbes, Behn, More and Defoe-as a borderline case-are significantly earlier texts, whereas the samples by Hobbes, Mill, and More all come from philosophical essays. As an early female playwright, Behn is also and understandably an outsider. Furthermore, clustering for gender seems to be quite successful. The family tree presents itself naturally in two parts, the upper part of which (from Defoe to Stoker) is predominantly populated by male authors (21 out of 30 or a score of 70%) and the lower part is strongly populated by female authors (16 out of 20 or a score of 80%). Since up to the end of the Victorian period, that is up to the beginning of the twentieth century, female authors are generally observed to adopt the prevailing male style of writing, the reason why four male authors (Kipling, James, Trollope, and Hardy) appear in the female part of the tree might be more interesting to study. In the second tree that shows the family tree after feature selection we can distinguish five groups of gender clusters with 11 exceptions (or 22%); six women writers (Stowe / Austin, Shelley / Ferber, Porter, Behn) and five male authors (Defoe / Collins, Trollope, James, Hardy). Aggregating the results from the first tree with the results from the gender-related stylistic family tree presented in Figure 2 reduces the initial female gender problem from 9 to 3 cases (only A. Brontë, Canfield, and, C. Brontë are correctly clustered within female groups after feature selection) and the male gender problem from 4 to 3 (James, Trollope, and Hardy). However, this clustering introduced two new problematic names: Defoe and Collins which, together with the remaining names, deserve further research.

5. Conclusions and Further Research

Without claiming any relevance for these particular family trees, it seems clear to us that specific literary style hypotheses can be tested using similar approaches. Close cooperation between literary scholars and computational linguists is essential for this.

We have shown that robust text analysis can bring a new set of tools to literary analysis. Specific hypotheses can be tested and new insights can be gained by representing the work (or different works) of authors as profiles and applying clustering and learning techniques to them. In future work we will investigate more specific literary hypotheses, and generalize the appoach to the analysis and comparison of individual books of authors rather than random samples of their work.

6. References

S. Argamon, M. Koppel, J. Fine, and A. Shimoni. 2003. Gender, genre, and writing style in formal written texts. *Text*, 23(3).

H. Baayen, H. Van Halteren, and F. Tweedie. 1996. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11(3):121–131.

S. Bucher-Gillmayr. 1996. A computer-aided quest for allusions to biblical text within lyric poetry. *Literary and Linguistic Computing*, 11(1):1–8.

J. Burrows. 1987. *Computation into criticism: A study of Jane Austen's novels and an experiment in method*. Oxford: Clarendon Press.

T. Corns. 1982. *The development of Milton's prose style*. Oxford: Clarendon Press.

H. Craig, 2004. *A Companion to Digital Humanities*, chapter Analysis and authorship studies, pages 273–288. Malden, MA/Oxford/Carlton, Victoria: Blackwell Publishing.

W. Daelemans and A. van den Bosch. 2005. *Memory-Based Language Processing*. Studies in Natural Language Processing. Cambridge, UK: Cambridge University Press.

J. Diederich, J. Kindermann, E. Leopold, and G. Paass. 2000. Authorship attribution with Support Vector Machines. *Applied Intelligence*, 19(1-2):109–123.

P. Fortier, 1989. *Literary Computing and Literary Criticism: Theoretical and Practical Essays on Theme and Rhetoric*, chapter Analysis of twentieth-century French prose fiction, pages 77–95. Philadelphia: University of Pennsylvania Press.

P. Fortier, 1996. *Research in Humanities Computing 5: Papers from the 1995 ACH-ALLC Conference*, chapter Categories, theories, and words in literary texts, pages 91–109. Oxford: Oxford University Press.

M. Gamon. 2004. Linguistic correlates of style: Authorship classification with deep linguistic analysis features. In *Proceedings of COLING 2004*, pages 611–617.

S. Hockey. 2000. *Electronic Texts in the Humanities*. Oxford: Oxford University Press.

D. Holmes. 1994. Authorship Attribution. *Computers and the Humanities*, 28(2):87–106.

N. Ide, 1986. *Méthodes quantitatives et informatiques dans l'étude des textes. In honour of C. Muller*, chapter Patterns of imagery in William Blake's 'The Four Zoas', pages 497–505. Geneva: Slatkine.

N. Ide, 1989. *Literary Computing and Literary Criticism: Theoretical and Practical Essays on Theme and Rhetoric*, chapter Meaning and method: computer-assisted analysis of Blake, pages 123–141. Philadelphia: University of Pennsylvania Press.

H. Ilsemann. 1995. Computerized drama analysis. *Literary and Linguistic Computing*, 10(1):11–21.

M. Julliard and X. Luong. 1997. Words in the hood. *Literary and Linguistic Computing*, 12(2):71–78.

M. Julliard and X. Luong. 2001. On consensus between tree-representation of linguistic data. *Literary and Linguistic Computing*, 16(1):59–76.

E. Kelih, G. Antíc, P. Grzybek, and E. Stadlober, 2005. *Classification of author and/or genre? The impact of word length*, pages 498–505. Heidelberg: Springer.

D. Khmelev and F. Tweedie. 2001. Using Markov chains for identification of writers. *Literary and Linguistic Computing*, 16(4):299–307.

M. Koppel, S. Argamon, and A. Shimoni. 2003. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412.

O. Kukushkina, A. Polikarpov, and D. Khmelev. 2001. Using literal and grammatical statistics for authorship attribution. *Problemy Peredachi Informatsii*, 37(2):96–108. Translated as Problems of Information Transmission.

C. Labbé and D. Labbé. to appear, 2006. A tool for literary studies: intertextual distance and tree classification. *Literary and Linguistic Computing*. Advance access: October 27, 2005.

I. Lancashire, 1993. *The Digital Word: Text-Based Computing in the Humanities*, chapter Computer-assisted critical analysis: a case study of Margaret Atwood's Handmaid's Tale, pages 291–318. Cambridge, MA/London: The MIT Press.

M. Marcus, B. Santorini, and M. Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics: Special Issue on Using Large Corpora*, 19(2):313–330.

W. McCarty. 2003. Depth, markup and modelling. *Text Technology*, 12(1).

R. Potter. 1981. Character definition through syntax: significant within-play variability in 21 English language plays. *Style*, 15:415–434.

R. Potter, 1989. *Literary Computing and Literary Criticism: Theoretical and Practical Essays on Theme and Rhetoric*, chapter Changes in Shaw's dramatic rhetoric, pages 225–258. Philadelphia: University of Pennsylvania Press.

J. Raben. 1965. A computer-aided study of literary influence: Milton to Shelley. In *Literary Data Processing Conference Proceedings*, pages 230–274. White Plains: IBM.

D. Robey. 2000. *Sound and Structure in the Divine Comedy*. Oxford: Oxford University Press.

T. Rommel, 2004. *A Companion to Digital Humanities*, chapter Literary Studies, pages 88–96. Malden, MA/Oxford/Carlton, Victoria: Blackwell Publishing.

F. Sebastiani. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

M. Spencer, B. Bordalejo, P. Robison, and C. Howe. 2003. How reliable is a stemma? An analysis of Chaucer's Miller Tale. *Literary and Linguistic Computing*, 18(4):407–422.

E. Stamatatos, N. Fakotakis, and G. Kokkinakis. 1999. Automatic authorship attribution. In *Proceedings of EACL 99*.

E. Stamatatos, N. Fakotakis, and G. Kokkinakis. 2001. Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35(2):193–214.

K. Steele, 1991. *A TACT Exemplar*, chapter 'The Whole Wealth of Thy Wit in an Instant'. TACT and the explicit structures of Shakespeare's plays, pages 15–35. Toronto: Centre for Computing in the Humanities.

R. Wisbey, 1971. *The Computer in Literary in Linguistic Research*, chapter Publications from an archive of computer-readable literary texts, pages 19–34. Cambridge: Cambridge University Press.

Female authors	Works	Number of Words	Period
Louisa-May Alcott	*Little Women*	100,000	1850+
Jane Austen	*Mansfield Park*	100,000	1800+
Mary Austin	*The Trail Book*	83,918	1900+
	The Land of Little Rain	16,082	
Aphra Behn	*The Rover*	75,673	1675+
	The City Heiress	24,327	
Anne Brontë	*The Tenant of Wildfell Hall*	100,000	1825+
Charlotte Brontë	*Jane Eyre*	100,000	1825+
Emily Brontë	*Wuthering Heights*	100,000	1825+
Frances Burnett	*The Secret Garden*	97,863	1900+
	A Little Princess	2,137	
Dorothy Canfield	*The Brimming Cup*	100,000	1900+
Willa Cather	*The Song of the Lark*	100,000	1900+
Agatha Christie	*The Secret Adversary*	95,852	1900+
	The Mysterious Affair at Styles	4,148	
Rebecca Davis	*Frances Waldeaux*	45,173	1875+
	Margret Howth	24,179	
	Life in the Iron-Mills	18,501	
	One Week an Editor	8,843	
	Walhalla	3,304	
Maria Edgeworth	*The Parent's Assistant*	100,000	1800+
George Eliot	*Silas Marner*	100,000	1875+
Edna Ferber	*Fanny Herself*	100,000	1900+
Mary Freeman	*The Heart's Highway*	85,980	1900+
	Copy-Cat and Other Stories	14,020	
Elizabeth Gaskell	*Sylvia's Lovers*	100,000	1850+
Charlotte Gilman	*What Diantha Did*	69,762	1900+
	Herland	30,238	
Susan Glaspell	*The Visioning*	100,000	1900+
Helen Jackson	*Ramona*	100,000	1875+
Eleanor Porter	*Just David*	100,000	1900+
Mary Shelley	*Frankenstein*	75,530	1800+
	Mathilda	24,470	
Harriet Stowe	*The Key to Uncle Tom's Cabin*	100,000	1850+
Edith Wharton	*The Age of Innocence*	100,000	1900+
Virginia Woolf	*Night and Day*	100,000	1900+

Table 1: Stylogene Literary Corpus: Female authors

Male authors	Works	Number of Words	Period
Jacob Abbott	*History of King Charles the Second of England*	65,076	1850+
	Aboriginal America	34,924	
Robert Browning	*Dramatic Romances*	57,541	1825+
	Sordello	42,459	
Wilkie Collins	*The Woman in White*	100,000	1850+
Charles Darwin	*The Voyage of the Beagle*	100,000	1900+
Daniel Defoe	*Moll Flanders*	100,000	1700+
Charles Dickens	*Dombey and Son*	100,000	1850+
Henry Fielding	*The History of Tom Jones, a Foundling*	100,000	1725+
Thomas Hardy	*Tess of the D'Urbervilles*	100,000	1875+
Nathaniel Hawthorne	*The Marble Faun*	100,000	1850+
Thomas Hobbes	*Leviathan*	100,000	1650+
Henry James	*The Portrait of a Lady*	100,000	1875+
James Joyce	*Ulysses*	100,000	1900+
Rudyard Kipling	*Actions and Reactions*	83,648	1900+
	Captains Curageous	16,352	
D.H. Lawrence	*Women in Love*	100,000	1900+
Herman Melville	*Moby Dick*	100,000	1850+
J.S. Mill	*On Liberty*	53,773	1850+
	The Subjection of Women	46,227	
Thomas More	*Dialogue of Comfort against Tribulation*	100,000	1525+
E.A. Poe	*A Descent into the Maelstrom*	100,000	1825+
	The Gold-Bug		
	Mellonta Tauta		
Laurence Sterne	*The Life and Opinions of Tristram Shandy*	100,000	1750+
Bram Stoker	*Dracula*	100,000	1900+
Jonathan Swift	*Gulliver's Travels*	100,000	1725+
Anthony Trollope	*Can You Forgive Her?*	100,000	1850+
Mark Twain	*The Innocents Abroad*	100,000	1850+
H.G. Wells	*The World Set Free*	73,522	1900+
	The War of the Worlds	26,478	
Oscar Wilde	*The Picture of Dorian Gray*	95,213	1875+
	Lord Arthur Savile's Crime	4,787	

Table 2: Stylogene Literary Corpus: Male authors