Measuring the style of chick lit and literature

Jautze, Kim
kim.jautze@huygens.knaw.nl
Huygens Institute for the History of the Netherlands

1. Introduction

Novelistic genres have certain formal conventions. By close reading novels of various genres, structural features such as the differences in theme, motifs and plot can be observed. According to Jockers (2013), there seem to be stylistic differences, caused by the author's linguistic choices, between genres as well. In this paper I examine the stylistic fingerprints of chick lit and "high brow" literature.[1] A previous study into the deep syntactic structures of the two genres showed that literary authors tend to use more complex sentences and employ a more descriptive language, whereas chick lit to a greater extent resembles colloquial speech (Jautze et al., 2013). In the current study I complement this syntactic characterization with a stylometric analysis of the function words. These words relate to syntactic structure because they add grammatical information by organizing and connecting the content words. The question arises if the most frequent function words (MFWs) differentiate between the two genres. And if so, do these linguistic patterns give more insight into the two genre styles?

2. Background

Stylometrists who study linguistic patterns in fiction typically focus on classification tasks, e.g. authorship attribution or text genre detection. The latter studies usually examine how well certain texts can be identified into pre-defined classifications; for instance *editorials*, *newspapers* and *literature* (Stamatatos et al., 2000). Stylometric studies of novelistic genres seem to be scarce.

One of few is performed by Louwerse et al. (2008). Predominantly, they focus on computationally discriminating literary from non-literary texts (newspapers and dialogue), but additionally find that the bigram 'and in' differentiates between Star Wars novels and quality literature. In a more recent study differences in the distribution of unigrams and parts of speech are found among eight prose genres (Ashok et al., 2013). One

[1] Chick-lit novels humorously address the challenges of young urban female protagonists.

of the most extensive studies into novelistic genres is performed by Matthew Jockers (2013) and his colleagues at the Stanford Literary Lab (Allison et al., 2011). They examine to what extent formal conventions can be detected at the level of the high-frequency function words of several nineteenth-century British genres. Jockers (2013) concludes that these genres to a certain degree have measurable linguistic fingerprints, and that linguistic decisions of authors are dependent upon their genre choices.

An interesting next step would be to analyze and interpret these linguistic fingerprints, as has been done for authorial markers by Burrows and Craig (2012). They apply the stylometric approach for stylistic research into authorial style by interpreting the differences in the use of function words. In this paper, I adopt their approach to examine the stylistic differences between chick lit and literature.

3. Materials and method

According to Jockers (2013), it is hard to distinguish linguistic fingerprints that are related to the time of writing from actual genre signals. This means that when one wants to examine genre fingerprints, the chronology factor must be ruled out as far as possible. My corpus therefore comprises 32 original Dutch novels (16 literary and 16 chick-lit) of the last two decades.[2]

In order to computationally examine the style of the two genres I start with a stylometric approach to search for the style markers. The stylo package in R compiles a word frequency list for the entire corpus (Eder et al., 2013). Statistical procedures such as Principal Components Analysis (PCA) and the Bootstrap Consensus Tree provide visual insight into the distances between texts and authors, and perhaps also between genres.

If the two can be distinguished, I want to explore the language patterns to characterize the two genre styles. In Jautze et al. (2013) we conclude that when analyzing fiction, one should take into account the differences of language use in descriptive passages and dialogues. Egbert (2012) argues something similar and shows that lexicogrammatical features can be captured in three dimensions of discourse presentation. Two of these dimensions I will adopt in this analysis in order to analyze the linguistic patterns: (i) description versus thought representation and (ii) dialogue versus narrative.

[2] This is the same corpus as has been studied in Jautze et al. (2013).

4. Results

Figure 1 shows that the authors as well as the two genres cluster together (the abbreviations indicate the pre-defined genres). This Bootstrap Consensus Tree is a mean of ten cluster analyses, varying from 100-1000 MFWs with an increment of 100. The corpus is culled at 100%, which means that words that are unique for individual texts are removed. It is striking that the chick-lit writers are more grouped together than the literary authors, which indicates that there is more variation within the literary writing style than within the style of the chick-lit writers.

Figure 1: A Bootstrap Consensus Tree showing average similarity of texts based on the frequencies of 100-1000 MFW.

In order to examine linguistic patterns behind this genre-distinction, the word frequencies are analyzed. A PCA uses the MFWs as variables according to which the texts are correlated in a matrix. Figure 2 shows that the 100 MFWs map the genres in separate areas of the graph, except for chick-lit writer Wilma Hollander. She sides with the literary authors.

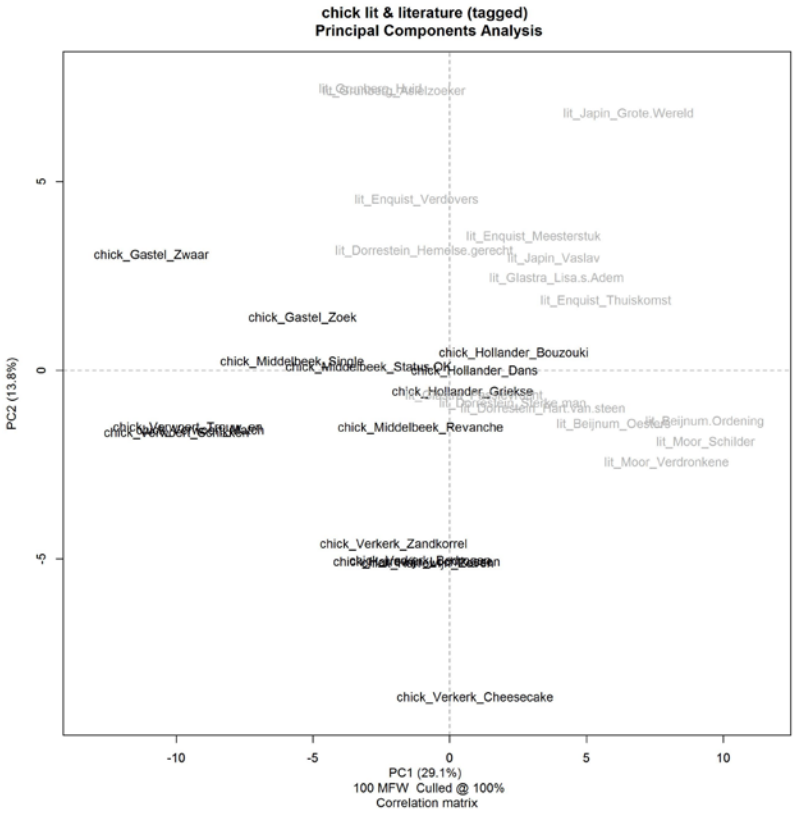

Figure 2: A PCA showing the plotting of texts based on the weightings of 100 MFW.

The two components of the PCA together account for 43,7% of the variation between the novels. The word-variables have their own weightings for each component according to which the texts are scored in the matrix (e.g. Figure 3). In the previous study by Jautze et al. (2013) the novels were parsed with the Alpino parser (Bouma et al., 2001). The parts-of-speech tags made it possible to separate homographic word forms, as in *zijn* ('his') and *zijn* ('are').

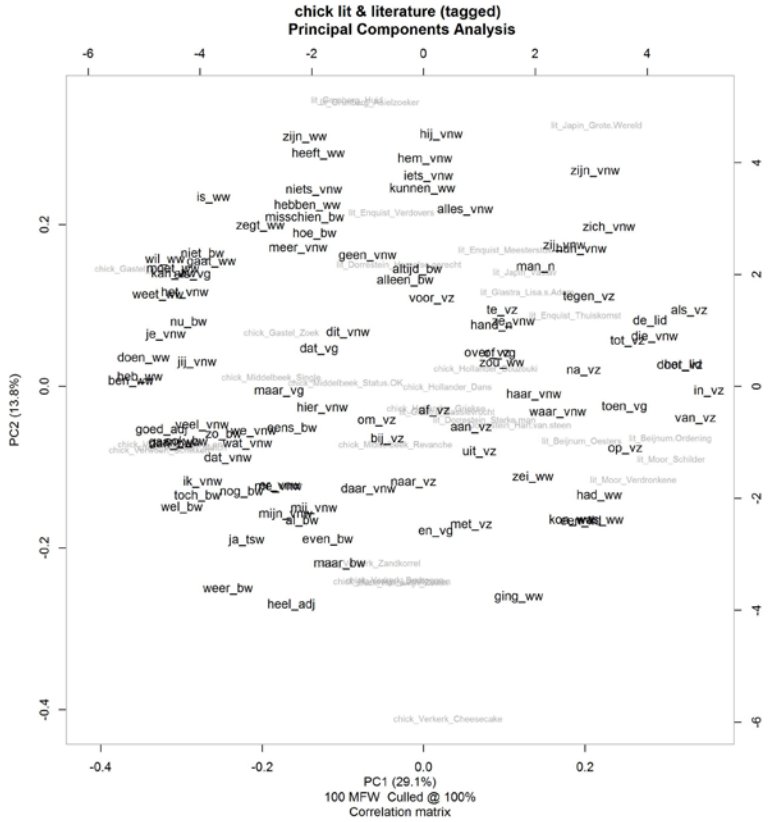

Figure 3: A PCA showing the plotting of 100 word-variable weightings.

POS tags	Translation tags
Vnw	Pronoun
Ww	Verb
Bw	Adverb
Vg	Conjunction
Vz	Preposition
Lid	Determiner
Adj	Adjective
N	Noun
Tsw	Interjection

Table 1: Translation POS tags

With regard to Egbert's dimensions, it can be argued that the literary authors employ more *descriptions* and *narratives*, whereas in chick lit more *thought representations* and *dialogues* are staged. Indicative for the descriptive dimension is the high amount of prepositions, the use of determiners and the demonstrative *die* ('that'). Prepositions express spatial or temporal relations between subjects and/or objects, and

therefore are used for detailed-oriented description. Along with the use of determiners and demonstratives, it indicates that the literary authors use relatively more nouns. These findings can be underlined by the results of Jautze et al. (2013), that show that noun phrases and prepositional phrases occur more frequently in the literary books than in the chick-lit novels.

Other frequent "literary" function words in the PCA are third person pronouns such as *hij* and *hun* ('he' and 'their/them'), indefinite pronouns such as *iets* and *alles* ('something' and 'everything') and verbs in the past tense. According to Egbert (2012), these linguistic features belong to the narrative dimension. Especially the past tense verbs indicate that the literary narrators describe events. The chick-lit writers on the other hand, employ more present tense verbs, and first and second person pronouns such as *ik*, *mij* and *jij* ('I', 'me' and 'you'). These, as well as the demonstratives *dat* and *daar* ('that' and 'there'), are argued to be indicative for the dialogue dimension.

Moreover, at the chick-lit side of the plot a lot of words are mapped that relate to the dimension of thought representation. Function words like the mental verb *weet* ('know'), the indefinite pronoun *veel* ('many'), the affective adjectives *heel* and *goed* ('very' and 'good'), the possibility modal *kan* ('can') and the likelihood adverb *misschien* ('maybe') offer insight into the narrator's or character's psyche. The chick-lit authors also employ certain adverbs (*maar*, *toch*) that can cause a certain emphatic effect. It could be compared with 'there are *only*/*like* seven'. It shows a character's or narrator's involvement, and it belongs to a more colloquial language register.

5. Conclusion

The results of this paper show that stylometric analysis can be applied for stylistic research of literary genres. The linguistic patterns detected in this small corpus suggest that the literary authors have a more detailed-orientated descriptive style when compared to the chick-lit style, that tends to be more informal and involved. The next step will be to evaluate these methods on a larger corpus of literary and chick-lit novels (perhaps by using translations), and to explore other literary genres.

Acknowledgements

This study is part of The Riddle of Literary Quality Project, supported by the Royal Netherlands Academy of Arts and Sciences as part of the Computational Humanities program.[3] I am grateful to my supervisor professor Karina van Dalen-

[3] In this project we explore the assumption if formal characteristics play a role in the aesthetic appreciation of novels. Cf. http://literaryquality.huygens.knaw.nl/

Oskam and to Corina Koolen and Andreas van Cranenburgh for reading my drafts, and to Andreas for assisting with tagging the parts of speech in the novels.

References

Allison, S., Heuser, R., Jockers, M., Moretti, F. and Witmore, M. (2011). Quantitative Formalism: An Experiment. In *Pamphlets of the Literary Lab* **1**, http://litlab.stanford.edu/LiteraryLabPamphlet1.pdf, (accessed on 24 October 2013).

Ashok, V.G. , Feng, S., and Choi, Y. (2013). Success with Style: Using Writing Style to Predict the Success of Novels. In *Empirical Methods on Natural Language Processing*, Seattle. 1753-1764, http://aclweb.org/anthology/D/D13/D13-1181.pdf, (accessed on 28 October 2013).

Bouma, G., Van Noord, G. and Malouf, R. (2001). Alpino: Wide-coverage computational analysis of Dutch. In *Language and Computers*, **37 (1)**. 45–59.

Burrows, J. and Craig, H. (2012). Authors and Characters. In *English Studies* **93 (3)**. 292-309.

Eder, M., Kestemont, M. and Rybicki, J. (2013). Stylometry with R: a suite of tools. In *Digital Humanities 2013: Conference Abstracts*. Lincoln (NE). 487-489.

Egbert, J. (2012). Style in nineteenth century fiction. A Multi-Dimensional analysis. In *Scientific Study of Literature* **2 (2)**. 167-198.

Jautze, K., Koolen, C., Van Cranenburgh, A. and De Jong, H. (2013). From high heels to weed attics: a syntactic investigation of chick lit and literature. In *Proceedings of the Workshop on Computational Linguistics for Literature*. Atlanta (GA). 72-81.

Jockers, M. L. (2013). *Macroanalysis: Digital Methods and Literary History*. Illinois: University of Illinois Press.

Louwerse, M., Benesh, M.N. and Zhang, B. (2008), Computationally discriminating literary from non-literary texts. In: Zyngier, S., Bortolussi, M., Chesnokova, A. & Auracher, J. (Eds.), *Directions in Empirical Literary Studies. Linguistic Approaches to Literature* **5**, Amsterdam. 175-191.

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2000). Text Genre Detection Using Common Word Frequencies. In *Proceedings of the 18th conference on Computational linguistics* **(2)**. Stroudsburg (PA). 808–814.