Stylometry applied to book preferences

Peter Boot, peter.boot@huygens.knaw.nl

Paper for DH Benelux 2017, https://dhbenelux2017.eu/wp-content/uploads/sites/187/2017/06/Abstracts_DHBenelux_Tuesday.pdf#page=66

Introduction

One of the oldest and most active fields in Digital Humanities is authorship attribution. It has been shown many times that writers have a characteristic style that can be used to tell them apart (e.g. Burrows, 2002). It is also well known that word usage can be used to predict personality characteristics (e.g. Noecker, Ryan, & Juola, 2013). Personality characteristics in turn are related to preferences in different art forms (e.g. Cantador, Fernández-Tobías, Bellogín, Kosinski, & Stillwell, 2013). This suggests that, as one would hope, the stylistic differences whereby we tell authors apart (such as differences in function word usage) are not just meaningless preferences for one function word over another, but are related to artistic preference, in a way that is still to be clarified.

This paper, continuing earlier work (Boot, 2014), tries to contribute to that clarification, in that it will remove the middle term (the personality characteristics) and show that there is a direct relation between the words that people use and their preferences in art, in this case, for books. The writers that I study here are the writers of book reviews, not books. In the first section, I will use book reviews and ratings from book discussion sites and show correlations between word usage and book ratings. In the second section, I will take an exploratory approach and create a clustering of reviewers by word usage. For the two clusters, I will then look at their preferred word usage, as well as the word usage in the book descriptions of their preferred books.

Correlations between word usage and ratings

The data that the paper uses were collected from a number of Dutch book discussion sites. These sites include hebban.nl, lezerstippenlezers.be, bol.com and the now defunct sites watleesjij.nu and dizzie.nl.

The correlations were computed as follows: I selected reviews from users who had written at least 100000 characters, excluding some users with multiple accounts. I computed relative word frequencies in their reviews, and normalized the results (center around zero and divide by the standard deviation). In order to remove words with thematic links to books (murder, war, castle, love) I limited the computation to words defined as function words in the Dutch LIWC 2007 dictionary (Boot, Zijlstra, & Geenen, 2017, in press). For the same users I retrieved the book ratings and created a matrix of users by rating, excluding books that were rated only once. I computed the bias corrected distance correlation (a multivariate generalization of the correlation coefficient, see Székely & Rizzo, 2013) between the two matrices, and repeated that computation for reviews in all genres, in literature and in the literary thriller. The results are given in the first row of Table 1.

To be absolutely sure that no content-aspects of the reviews were reflected in the word usage, I repeated the computation using Part-of-speech-tags . The texts were tagged using Treetagger and instead of the relative word frequencies I used relative frequencies of POS bigrams. The results are given in the second row of the table.

Table 1

Correlations with p-values	**All genres** 189 reviewers 166 reviews (avg.)	**Literature** 41 reviewers 126 reviews (avg.)	**Literary thriller** 32 reviewers 88 reviews (avg.)
function words (200) vs. ratings	0.20 (0.000)	0.16 (0.000)	0.41 (0.000)
POS bigrams (100) vs. ratings	0.16 (0.000)	0.10 (0.002)	0.22 (0.000)

It is hard to interpret these correlation sizes, but it is clear that there are very significant correlations between function word usage and book ratings. The fact that these correlations persist even when looking at POS bigrams shows that the relation is to some extent based purely on linguistic style, not on content. Why sequences of POS-tags should be related to literary preference is an intriguing question that this paper will not solve.

Exploratory analysis

To get a feel for what this correlation might mean in terms of real reviews and ratings, I created a clustering based on function word usage for a group of reviewers. I removed a few outliers and was left with two clusters, cluster 1 containing 20 reviewers and cluster 2 containing 11.

I then looked at their reviews and preferred books. A sample of reviews from cluster 1 showed their informal, direct and very personal writing, characteristics that were much less prominent in cluster 2. This impression is confirmed when looking at contrastive keywords in the reviews of both clusters. The 20 key words with the largest effect size (Gabrielatos & Marchi, 2011) for both clusters are shown in table 2. It is clear cluster 1 prefers the first person, cluster 2 has more interest in writing.

Table 2

Cluster	Preferred review words
1	thought (was of the opinion), very, because, completely, me, actually, therefore, read (past part.) , beautiful, after all, had, have (1st pers. sing.), am, I, very, all, good, otherwise, yet, again
2	writer (fem.), writer, novel, reader, years, under, know, these, characters, one, between, gives, second, the, them, of, until, end, in, who

Turning to the ratings, while there were many books that were rated significantly higher by one of the groups, the preferences were hard to understand in terms of taste. Ratings summed by genre didn't show a very clear picture either. It was only when looking at contrastive word usage in the (publisher-provided) book descriptions for books read by either cluster that a clearer picture emerged.

Table 3

Cluster	Key words in preferred book descriptions
1	thriller, investigation, police, murdered, murder, case, body, someone, further, secret, above, know, very, sits, very, disappeared, within, nothing, appears, found, become,

	part, truth, books, there, something, else
2	in which, without, about, parents, family, city, big stories, last, exist, us, we, writer, history, love, country, tells, century, novel, Netherlands, war

Here it becomes clear that cluster 1 prefers thrillers and police novels, while cluster 2 has a less-focussed interest in family, writing and the country. It is worthwhile to repeat that these clusters of content words result from clustering reviewers on the basis of function words.

Conclusion

Taken together, the correlations and the exploratory analysis show that there is a relation between the function words that people use and their preferences for books. This relation still holds at the level of part-of-speech tags. This clearly shows that the word usage that helps tell authors apart is to some extent related to artistic preference. A possible explanation would be that the reviewers unconsciously imitate the books they read in their use of function words. That seems unlikely, among other reasons because the effect is also visible when we just look at the reviews in a single genre (second and third column of table 1). The more likely explanation is that function word usage is at least in part determined by artistic preference and related personality characteristics. The 'fingerprint' metaphor that is often used in this context, with its suggestion of an essentially random identifier, unlikely to be related to artistic preference, must therefore be considered as inappropriate.

Literature

Boot, P. (2014). *Dimensions of literary appreciation. Word use and ratings on a book discussion site*. Digital Humanities 2014. Retrieved from http://dharchive.org/paper/DH2014/Paper-825.xml

Boot, P., Zijlstra, H., & Geenen, R. (2017, in press). The Dutch translation of the Linguistic Inquiry and Word Count (LIWC) 2007 dictionary. *Dutch Journal of Applied Linguistics, 6*(1).

Burrows, J. (2002). 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing, 17*(3), 267-287.

Cantador, I., Fernández-Tobías, I., Bellogín, A., Kosinski, M., & Stillwell, D. (2013). *Relating Personality Types with User Preferences in Multiple Entertainment Domains.* Proceedings of the 1st Workshop on Emotions and Personality in Personalized Services (EMPIRE 2013), at the 21st Conference on User Modeling, Adaptation and Personalization (UMAP 2013).

Gabrielatos, C., & Marchi, A. (2011). Keyness: Matching metrics to definitions. *Theoretical-methodological challenges in corpus approaches to discourse studies-and some ways of addressing them*.

Noecker, J., Ryan, M., & Juola, P. (2013). Psychological profiling through textual analysis. *Literary and Linguistic Computing, 28*(3), 382-387.

Székely, G. J., & Rizzo, M. L. (2013). The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis, 117*, 193-213.