Dimensions of literary appreciation. Word use and ratings on a book discussion site

Boot, P.

published in
Digital Humanities 2014
2014

document version
Publisher's PDF, also known as Version of record

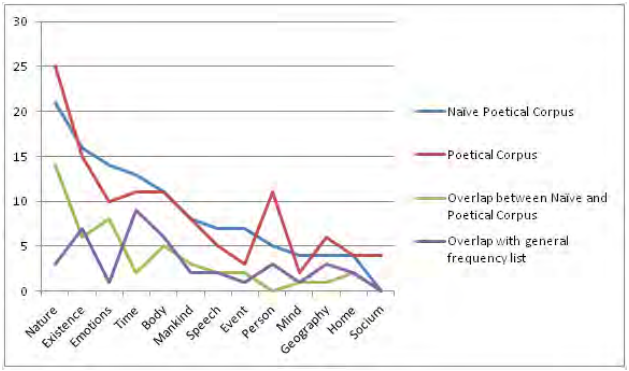
document license
CC BY

with the high poetical list (*problem*). The gap between naïve and poetic frequency list signals that there are some semantic zones where naïve poetry seems to be independent from the classical poetic canon. This trend can be defined as a self-actualization strategy which is in some sense opposite to the imitative strategy.

3. We took top 100 nouns of every list and compared their lexical distribution . The nouns had been grouped into abstract semantic domains. Some words could be associated with several domains due to their polysemy. As a result we have identified 13 semantic domains, 12 of them are shared between naïve poetry, high poetry and common frequency lists and the 13th is not presented in the naïve poetry list. The domains we have defined are as follows:

Mankind (everything that may characterize a person: *soul, beauty, name, heart, strength* etc.), Body, Emotions, Mind, Existence (*God, world, truth, time, fate* etc.), Speech, Person (*father, son, friend, enemy* etc.), Event (*love, happiness, past, disaster* etc.), Time, Nature, Geography (*road, hill* etc.), Home (*window, door* etc.). The 13th domain which is found only in the high poetical list and in the general frequency list is Social and it includes such words as people, labor, fame in the poetical list, and state, money, law etc. in the general list.

Analysis of the overlaps and varieties of the naïve and high poetical lists showed differences in the elaboration of the domains in two corpora. The general frequency list helped to draw out the words that are commonly frequent and their presence in the list cannot be understood as a signal of the poetic concentration on the domain. The results are demonstrated on the graph below:



As we can see from the graph, there are three zones of the naïve poetical sample that demonstrate high lexical variety of frequent nouns in comparison to the poetical corpus. These are Emotions, Event and Speech. Most of the words of those domains are not frequent in general lexicon. The lexical multiplicity can be explained by extensive strategy: the naïve poets do not use sophisticated verbal apparatus to express the conceptual space of the verse, but prefer straightforward lexical naming (*pain, wish, encounter, grief, love, question, answer* etc.)

References

Grishina E., Korchagin K., Plungian V., Sitchinava D. *Poeticheskij korpus v ramkah NKRYa: obschaja struktura i perspektivy ispol'zovanija* 'Natsional'nyj korpus russkogo jazyka: 2006-2008. Novye rezul'taty I perspektivy'. Saint Petersburg, 2009. P. 71-113. [Poetic Corpus in RNC: general structure and using perspectives]

Moretti, Franco. *Graphs, Maps, Trees: Abstract models for a literary history*. Verso, 2005.

Moretti, Franco. *Distant Reading*. Verso, 2013

Jockers, Matthew L. *Macroanalysis*. University of Illinois Press, 2013

Dimensions of literary appreciation. Word use and ratings on a book discussion site

Boot, Peter
peter.boot@huygens.knaw.nl
Huygens ING, Netherlands, The

Introduction

The appreciation of literature is a subjective process. In reading and judging books, characteristics of individual readers interact with characteristics of books and their reputation. This paper looks at book ratings on a book discussion site and tries to assess the role of individual readers' characteristics in these ratings. For that purpose, the paper inspects on the one hand the textual properties of the review texts that readers contribute to this site, and on the other hand the ratings that they assign.

Given the well-established connection between word use frequencies and authorial style (e.g. Burrows, 2002; Burrows, 2003), the paper hypothesizes that these same style markers in texts by readers will correlate with these readers' quality judgments about books. Patterns in word usage are known to reflect aspects of readers' psychological make-up (Argamon et al., 2005; Noecker et al., 2013; Pennebaker et al., 2003), and these psychological properties, e.g. the Big Five personality dimensions, are related to aesthetic preferences in many fields (Golbeck and Norris, 2013; Gridley, 2013; Zweigenhaft, 2008), including books and literature (Cantador et al., 2013; Wiersema et al., 2012).

Aesthetic appreciation has been shown to be a multi-faceted process (Myszkowski et al., 2014; Rentfrow et al., 2011). Here, I assume that literary appreciation is influenced by multiple aspects of the reader's psychology, such as, among others, his/her cognitive, affective and social dispositions. Therefore, besides investigating the over-all most frequently used words, as stylometry often does, I will also look at the high frequency words within the categories of cognitive, affective, and social words, as defined by LIWC (Pennebaker et al., 2007). I expect that the relative frequencies of e.g. individual social words (rather than the category frequencies that LIWC-based research typically uses) will capture to some extent the nature of a person's sociability and will to that extent also reflect how that sociability affects literary preferences.

Data

The data for this paper come from Dutch book discussion site watleesjij.nu (whatareyoureading.now). The site is similar to e.g. Goodreads, LibraryThing or lovelybooks.de: users rate, label and review books, they can evaluate reviews by others, can strike up friendships with and send messages to other users. I downloaded the site's content in June 2013. I investigate review texts and ratings contributed by the top 20 (in terms of total review length) contributors to the site (I removed two accounts that seemed to be used by multiple persons.) For each of these users, I create a file containing all of the review texts this user has contributed to the site. The average word count is 44036. I also collect the ratings (in terms of one to five stars) for all of the 624 books that were rated by at least two of the twenty users.

Method and results

As a first step, I compute correlations between the word use frequencies in each of the word categories and the book ratings. The word frequencies are represented as a matrix of zscores, where users are rows and words are columns. For the computation of the zscores I use Eder and Rybicki's stylo script (2011), then select only those words that form part of the relevant LIWC category. The ratings are given in a matrix with users as rows and books as columns. Non-rated books

are represented by 0. To assess the correlation between these matrices I rely on the (bias corrected) distance correlation and the associated significance test as described by Székely and Rizzo (2013). Table 1 reports the results, including the number of words that gave the best results for each category (However, for all categories except Affect the correlations were significant at the .01-level even for the top 25 words.) The table also gives the percentage of words belonging to the category in the review files.

Table 1. Bias-corrected distance correlation between word usage and book appreciation for different word categories

Category	bias corrected distance correlation	p-value	Optimum number of most frequent words	percentage of words in category
All words	.49	<.0001	2900	100
Function words	.36	<.0001	250	50
Affect words	.21	.0025	225	3
Cognitive words	.35	<.0001	375	18
Social words	.47	<.0001	125	12

The table shows that frequencies in each of the word categories are quite significantly correlated with the word ratings. The relatively low effect from affective words may be due to the low percentage of affect words in the texts.

The most striking result is no doubt the performance of the category of social words. In order to further investigate this effect, users were clustered in two groups, based on their usage of social words (I employed the pam partitioning function in R.) I then looked at contrastive word use of these clusters and at the books liked by the cluster members. The oppose function from Eder and Rybicki (2011) was used to find words preferred by either cluster. The results are given in table 2. The first cluster shows an interest in people and especially family that the second cluster, with its mostly cognitive or procedural interest, seems to lack entirely.

Table 2. Words preferred (from all words) by the two clusters (translated from Dutch). For cluster one, only the top 20 preferred words are given

Cluster	Preferred words
1	daughter, parents, family [nuclear], mother, woman, together, father, children, past, young, child, debut, house, brother, women, tells, love, marriage, family [extended], care
2	so, perhaps, page, of course, pity, well, read, precisely, actually, just, immediately, think, for instance, part, viz., believe, even, sort of, interesting, by the way

In order to find out the sort of books preferred by the clusters, I summed the book ratings by cluster. I then selected and diagrammed a subset of books, consisting of the ten books best liked by either cluster, the ten books best liked 'contrastively' by either cluster (computed by subtracting the ratings for cluster 1 from those for cluster 2), and the ten books best liked by both. After removal of duplicates, thirty books remained. Figure 1 displays the books with their ratings by the two clusters. Point and title size reflect popularity on the site. Grayscale represents genre. Point positions were slightly changed to avoid overlap. Lines between title labels and points were suppressed in the interest of clarity.

Fig. 1: Books as rated by the two clusters.

The figure seems to show some systematic differences in the preferences of the two clusters. Cluster 1, that uses mostly family-oriented words, seems to prefer slightly more popular books (larger point size). Cluster 2, that uses procedural or cognitive words, has strong preferences for a number of staunchly 'literary' works, such as those by Grass, Binet and classical Dutch authors. As to a potential preference for suspense novels, this figure does not allow us to draw any firm conclusions.

Discussion

The reason why different people prefer different books has often been sought in differing literary norms (e.g. Von Heydebrand and Winko, 2008). This explanation is not quite satisfactory, for two reasons: first because it does not explain why people develop different norms, and second because there are no a priori reasons why norms rather than, say pleasure or 'thrills' (Konecni, 2005) should determine one's preference for one book over another. This paper takes another approach and the results presented here tentatively establish the existence of a correlation between book preferences and patterns of word usage in several psychologically meaningful categories. Especially the relation between the pattern of usage of social words and literary appreciation seems very strong, confirming the importance of extraversion for aesthetic judgment noted by Furnham and Chamorro-Premuzic (2004), but appreciation is also clearly related to usage of cognitive words and of function words.

There are some obvious limitations to this experiment. The number of subjects is very small (dictated by the need to have a sufficient number of words). It would also have been better to use texts from another domain. However, an exploratory analysis of the effect of clustering based on social word usage seems to show that the verbally more 'social' group prefers the less literary or more popular novel. Given the small numbers, more than provisional results should perhaps not be expected.

Next steps should include clustering on the basis of other word categories, an investigation into the independent effect of these categories, and case studies at the level of individual readers. It would also be very interesting to see to what extent the literary norms that readers formulate in the reviews can be shown to be related to the word usage patterns as discussed here.

References

Argamon, S., S. Dhawle, M. Koppel and J.W. Pennebaker. (2005) *'Lexical predictors of personality type'*, Proceedings of the Joint Annual Meeting of the Interface and the Classification Society of North America.

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

Burrows, J. (2003) *'Questions of Authorship: Attribution and Beyond. A Lecture Delivered on the Occasion of the Roberto Busa Award ACH-ALLC 2001, New York'*. Computers and the Humanities **37**(1): 5-32.

Cantador, I., I. Fernández-Tobías, A. Bellogín, M. Kosinski and D. Stillwell. (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).

Eder, M. and J. Rybicki. (2011) *'Stylometry with R'*. Paper presented at Digital Humanities 2011: Conference Abstracts, Stanford University, Stanford, CA.

Furnham, A. and T. Chamorro-Premuzic. (2004) '*Personality, intelligence, and art*'.Personality and Individual Differences **36**(3): 705-715.

Golbeck, J. and E. Norris. (2013) 'Personality, movie preferences, and recommendations', Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining: ACM, pp. 1414-1415.

Gridley, M.C. (2013) '*Preference for Abstract Art According to Thinking Styles and Personality'*. North American Journal of Psychology **15**(3).

Konecni, V.J. (2005) *'The aesthetic trinity: Awe, being moved, thrills'*.Bulletin of Psychology and the Arts **5**(2): 27-44.

Myszkowski, N., M. Storme, F. Zenasni and T. Lubart. (2014) *'Is visual aesthetic sensitivity independent from intelligence, personality and creativity?'*. Personality and Individual Differences **59**(16-20.

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

Pennebaker, J.W., R.J. Booth and M.E. Francis. (2007) *'Linguistic Inquiry and Word Count (LIWC2007)'*, Linguistic Inquiry and Word Count (LIWC2007). Austin, TX.

Pennebaker, J.W., M.R. Mehl and K.G. Niederhoffer. (2003) *'Psychological aspects of natural language use: Our words, our selves'*. Annual review of psychology **54**(1): 547-577.

Rentfrow, P.J., L.R. Goldberg and R. Zilca. (2011) *'Listening, watching, and reading: The structure and correlates of entertainment preferences'*.Journal of personality **79**(2): 223-258.

Székely, G.J. and M.L. Rizzo. (2013) *'The distance correlation t-test of independence in high dimension'*. Journal of Multivariate Analysis **117**(193-213.

Von Heydebrand, R. and S. Winko. (2008) *'The qualities of literatures'*, The Quality of Literature: Linguistic Studies in Literary Evaluation. Amsterdam: Benjamins, pp. 223-239.

Wiersema, D.V., J. Van Der Schalk and G.A. van Kleef. (2012) *'Who's afraid of red, yellow, and blue? Need for cognitive closure predicts aesthetic preferences'*. Psychology of Aesthetics, Creativity, and the Arts **6**(2): 168.

Zweigenhaft, R.L. (2008) *'A do re mi encore: A closer look at the personality correlates of music preferences'*. Journal of individual differences **29**(1): 45.

An XML Schema to Interpret Networked Biographies: Reading Mid-Range

Booth, Alison
University of Virginia, United States of America

Martin, Worthy
University of Virginia, United States of America

Collective Biographies of Women, is an open-access project supported by the Institute for Advanced Technology in the Humanities, Scholars' Lab, and the English Department at the University of Virginia, as well as an ACLS Digital Innovation Fellowship. In recent years it has grown from an online bibliography of all English- language books that collect three or more short biographies of women into a digital prosopography that interrelates women, printed books, and narratives in what we call documentary social networks (introduced at DH 2013). CBW stands out as a literary study of prosopographies in the print era, and primarily the transatlantic nineteenth century (see the bibliography, http://womensbios.lib.virginia.edu). Most research that employs the term *prosopography* allies itself with history or classical and medieval studies, and today, relies on databases and websites. We work with the concept as it is often defined, as collective biography, that is, printed prose collections of short biographies (see the selective bibliography for a context on prosopography, nonfiction narrative, and our method of mid-range reading).

The CBW database associates some 8700 persons, 13,000 chapters (biographies), and more than 1200 books of various types published in English 1830-1940 (see developing database at http://cbw.iath.virginia.edu/cbw_db). Our project, however, is neither a textual archive nor a biographical database but an experiment in interpretation using the tools of DH to recognize the conventions of a genre, biography, and the history of gender conventions in a certain social context. Specifically, we want to get at the conditions of nonfiction, which generate multiple versions and cut and paste with relatively little respect for authorship. Could narrative theory of nonfiction be developed through a technique of digital markup that allows us to compare multiple versions of one life and interrelated types of person and text? With Daniel Pitti, Suzanne Keen, postdoctoral Project Manager Rennie Mapp, and teams of graduate assistants, we have developed and deployed a stand-aside XML schema, Biographical Elements and Structure Schema (BESS), in sample archives of digitized collective biographies that include designated individuals (e.g. all collections in our bibliography that include Caroline Herschel, the astronomer).

Briefly, BESS is an XML schema with a controlled vocabulary for narrative elements that appear in a given text:

- **StageofLife:** ***before, beginning, middle, culmination, end, after,*** *relative to the lifetimeofthebiography'ssubject*
- **EventType** ***e.g. illness, persona's***
  - **AgentType** ***e.g. mother, unnamed***
  - **Setting:**
    - **Location,** ***e.g. city***
    - **Structure,** ***e.g. school***
  - **Time:** Dates, TimeofDay, Season
- **PersonaDescription** ***e.g. physically daring***
- **Discourse:** ***e.g. retrospective, figureOrImage flower***
- **Topos:** ***e.g. influence, disgrace***

Each editor in a trained team creates a separate XML file that in effect is an annotated outline, tagging types of elements identified in numbered paragraphs of a TEI file of the biographical narrative (from 3-100+ paragraphs).