Aspects of Lexical Diversities

Introduction

Literary studies usually assume that literary texts have different levels of complexity and that the complexity of the language used, the richness or diversity of the vocabulary, is one of many determining factors (for example Koschorke 2016, Nan Da 2019). In other disciplines like corpus linguistics or second language acquisition research, there has been an interest in measurements for the vocabulary of language learners and how to determine whether a text is adequate for a particular skill level. In all of these fields, measures have been proposed over time, but only recently, attempts have been made to consolidate the diverse research into a comprehensive overview (for example Jarvis 2013; important forerunner on one dimension Tweedie/Baayen 1998).

In this paper, we propose a multi-dimensional model of vocabulary complexity. We provide a definition for each of the dimensions and an operationalization for most. We validate our operationalizations by comparing a collection of texts for adults with a collection of comparable texts for children. At the end, we show the usefulness of our proposed measure by applying it to literary texts.

Corpora

The validation corpora contain German non-fiction text from the educational magazine “Geo” (www.geo.de), a publication conceptually comparable to the “National Geographic”, and its offshoot for children called “Geolino”. For literary texts, we compare highbrow novels[1] with so called dime novels, a type of literary fiction mass-produced in long-lasting series and sold in kiosks rather than book stores.

[1] Novels by authors nominated for literature prizes

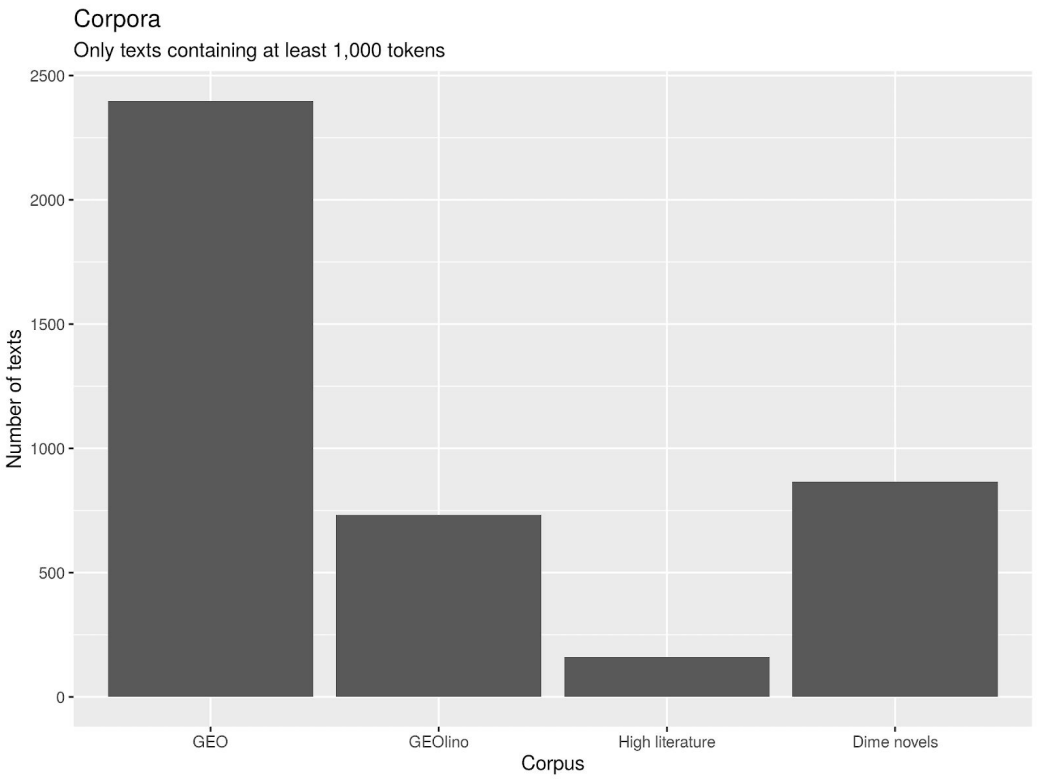

Fig. 1: Number of texts per collection

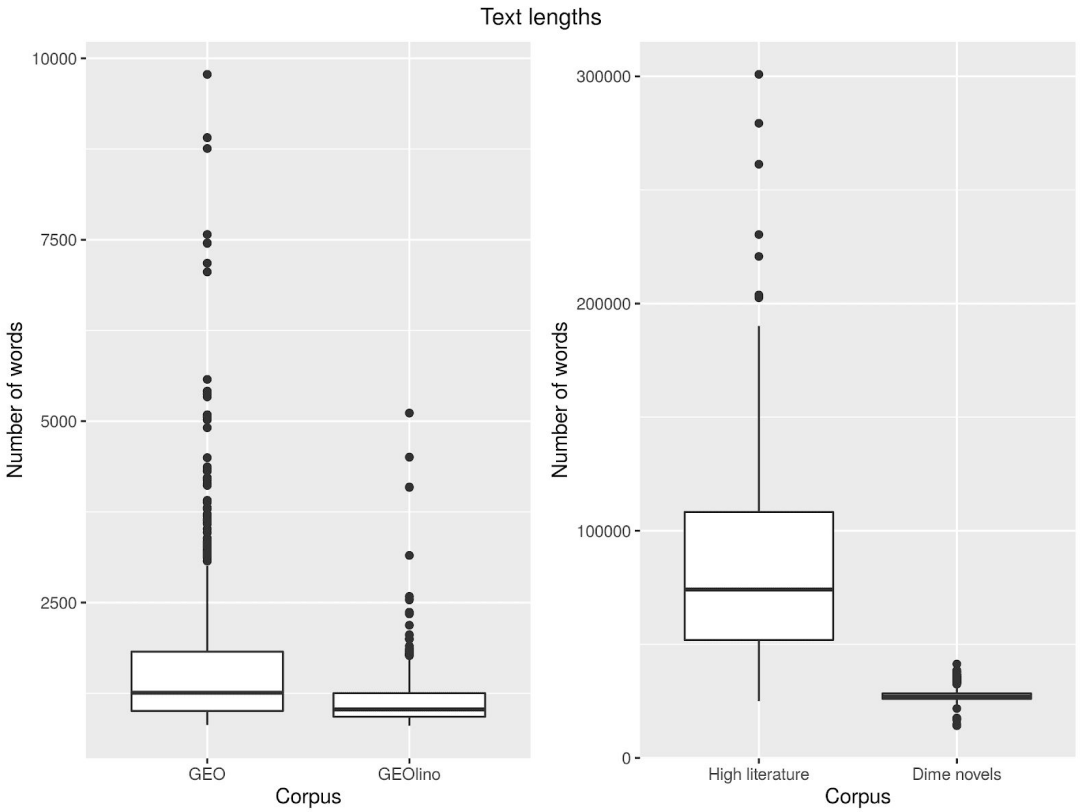

Fig. 2: Text length in collections

Aspects of complexity and measurement

Quantifying diversity is no trivial task. As Jarvis (2013b) points out, existing measures of lexical diversity often lack an underlying construct definition and intuitive concepts of diversity vary. Jarvis proposes six dimensions of lexical diversity to properly define the construct: variability, volume (which we do not consider separately), evenness, rarity, dispersion, and disparity. Additionally, we look at innovation, surprise, and density.

Variability

The most intuitive indicator of lexical diversity is the variability of the words used in a text. The most widely known measure is the type-token ratio (TTR) which you get by dividing the number of different words in a text (types) by the number of all words (tokens).
TTR depends heavily on sample size (as text volume increases, TTR decreases), making comparisons between different texts difficult. This has led to the development of many new measures, but no definitive solution has been found.
By computing these measures for text windows of the same size and averaging over all windows, we find that they fall into three groups which do not appear to measure the same thing: variations of TTR, measures which use only part of the frequency spectrum, and measures which use the whole frequency spectrum.

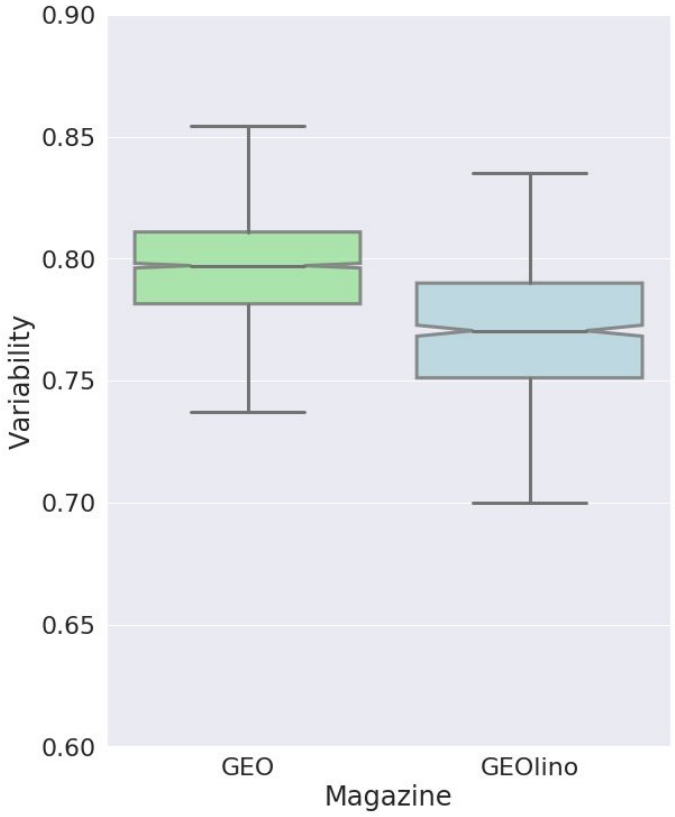

Fig. 3: STTR for GEO and GEOlino

Evenness

Evenness measures how evenly tokens are distributed among the different types. In biodiversity research, species evenness is usually measured as a biodiversity index proportional to its maximum, the most popular choice choice being the Pielou Evenness Index, calculated as Shannon entropy divided by its potential maximum value (Pielou 1966).

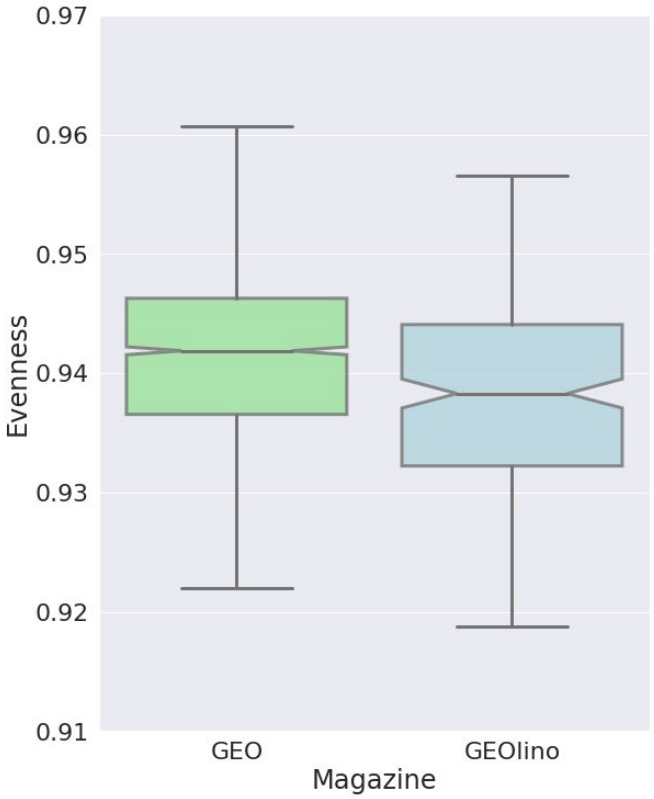

Fig. 4: *Evenness* for GEO and GEOlino

Rarity

A text containing many rare words will generally be perceived as more difficult and more complex than a text with a higher proportion of common or very frequent words.

We use a simple approach to model rarity. For each text, we compute the proportion of the 5,000 most frequent content words from the DECOW16BX corpus (Schäfer and Bildhauer 2012, Schäfer 2015), a large web corpus that covers many different registers.

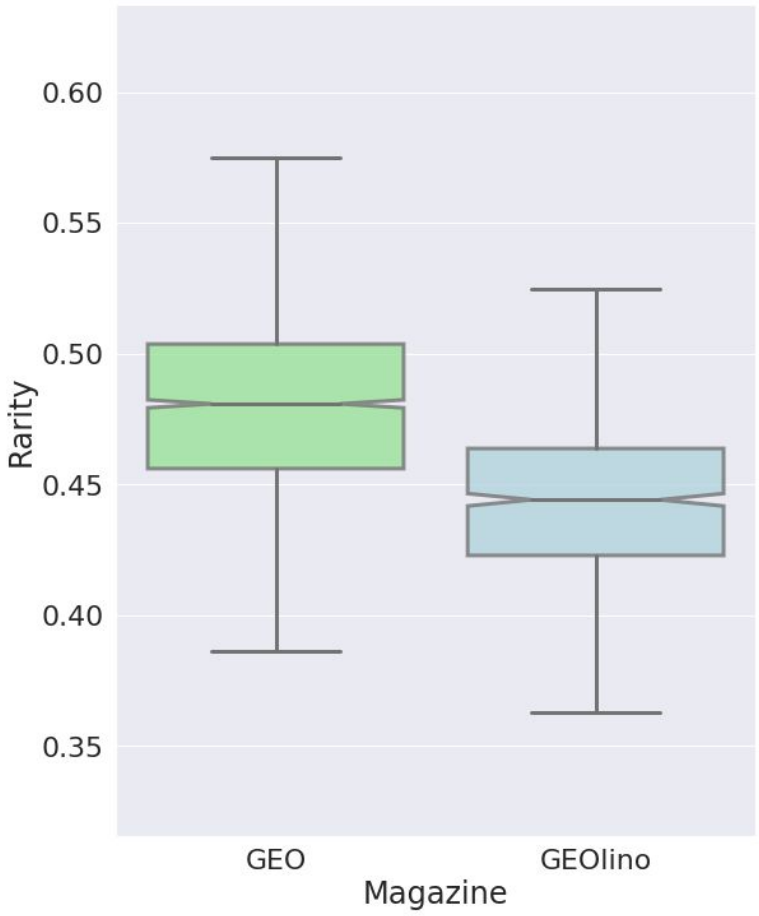

Fig. 5: *Rarity* for GEO and GEOlino

Dispersion

According to Jarvis (2013b), the perceived lexical diversity is higher if the occurrences of a particular type are more dispersed, whereas a more clustered pattern results in an impression of redundancy. To measure this effect, we use a window-based approach similar to the variability methods. Inside of these windows we calculate the zeta-diversity (Latcombe and McGeosh 2017) of types. Zeta-diversity is the average number of types occurring in 2,3, ..., n of these nested windows.

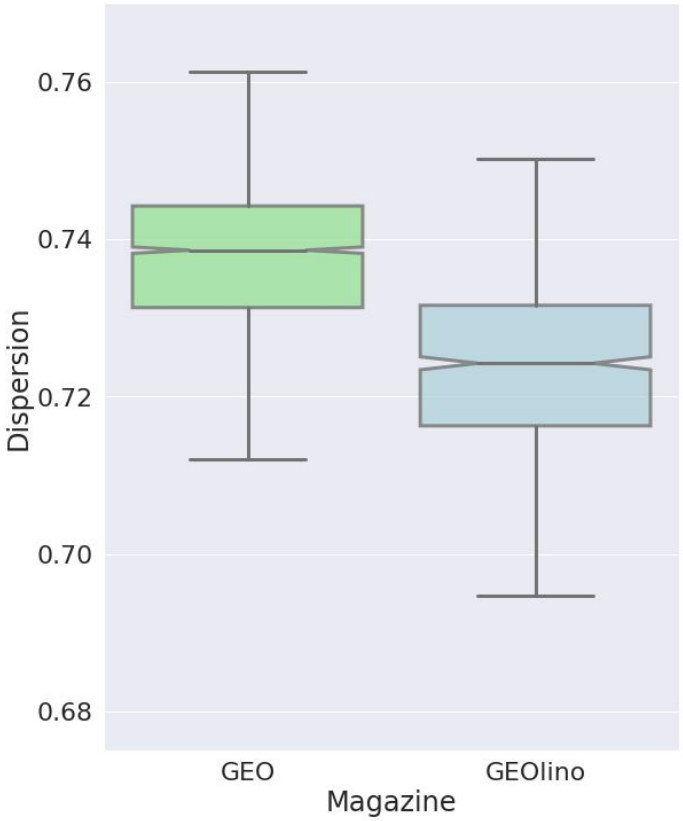

Fig. 6: Dispersion in GEO and GEOlino

Disparity

Lexical disparity follows the intuition that repetition is not completely covered by counting identical types, but needs to consider the occurrence of *similar* words on a semantic level as well. We model this concept following Cha et. al. (2017), using word embeddings and clustering to create a representation of a text segment and measure the distances to every other segment's representation.

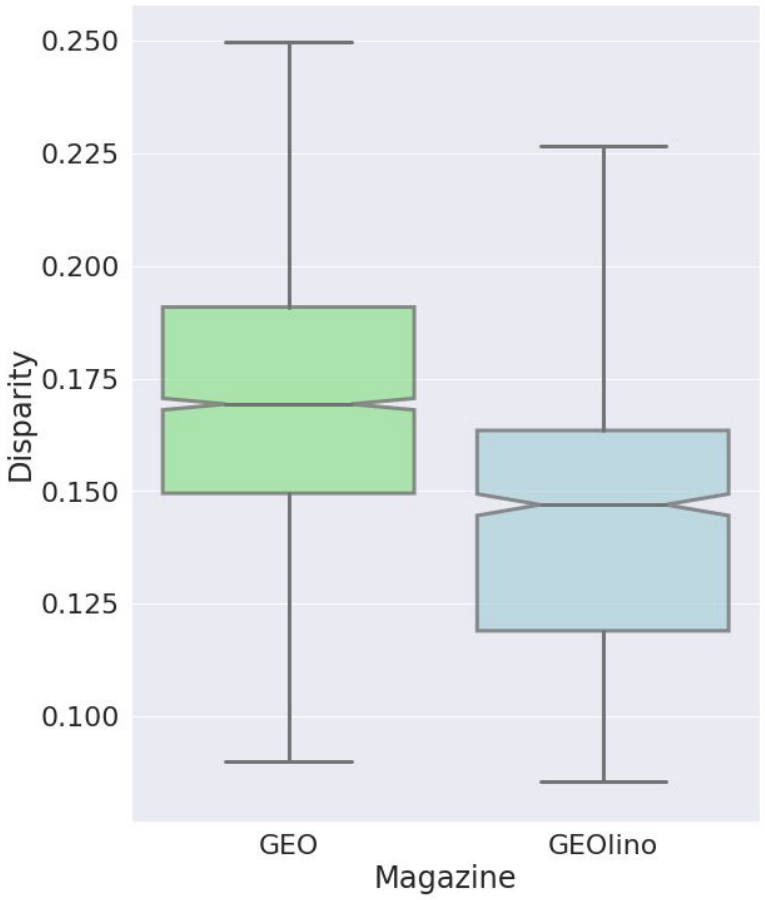

Fig. 7: Disparity for GEO and GEOlino

Surprise
We refer to surprise as the expectability or regularity of a text and the associated processing effort of a reader. We suggest to measure this property by the prediction quality of n-gram smoothing (Ney et al. 1994) or language models (Peters et. a. 2018). Not implemented because corpus balancing is a severe problem..

Innovation
We define innovation as creation of neologisms. The detection of new word is not trivial and has been in focus of various studies in the past (Falk et al. 2014, Klosa and Lungen 2018). At the moment this measure is not implemented.

Density
A text containing a higher proportion of content words (attributive, adverbial and predicative adjectives, nouns, proper nouns, and full verbs) can be considered more dense and thus, more complex. With a POS-tagged corpus, lexical density is easy to compute.

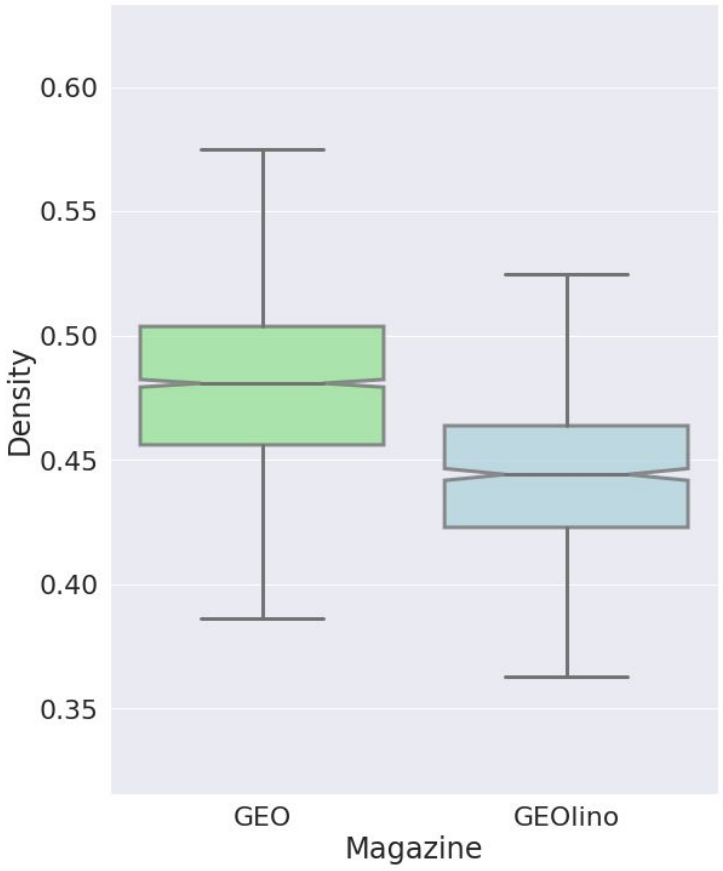

Fig. 8: Density for GEO and GEOlino

Application to Literature

Fig. 9 shows the cumulated complexity measures in the large bars. Counter to our expectations, our measures show science fiction to be more complex than high literature, but a look at the dimensions shows that mainly evenness, density and rarity are contributing to this. So we can see that we have different forms of lexical complexity at work here: In science fiction, a noun-heavy prose is depicting new worlds with new words. In high literature on the other hand, high variability, dispersion and rarity probably show the influence of a stylistic ideal which aims to avoid repetition and show elegance.

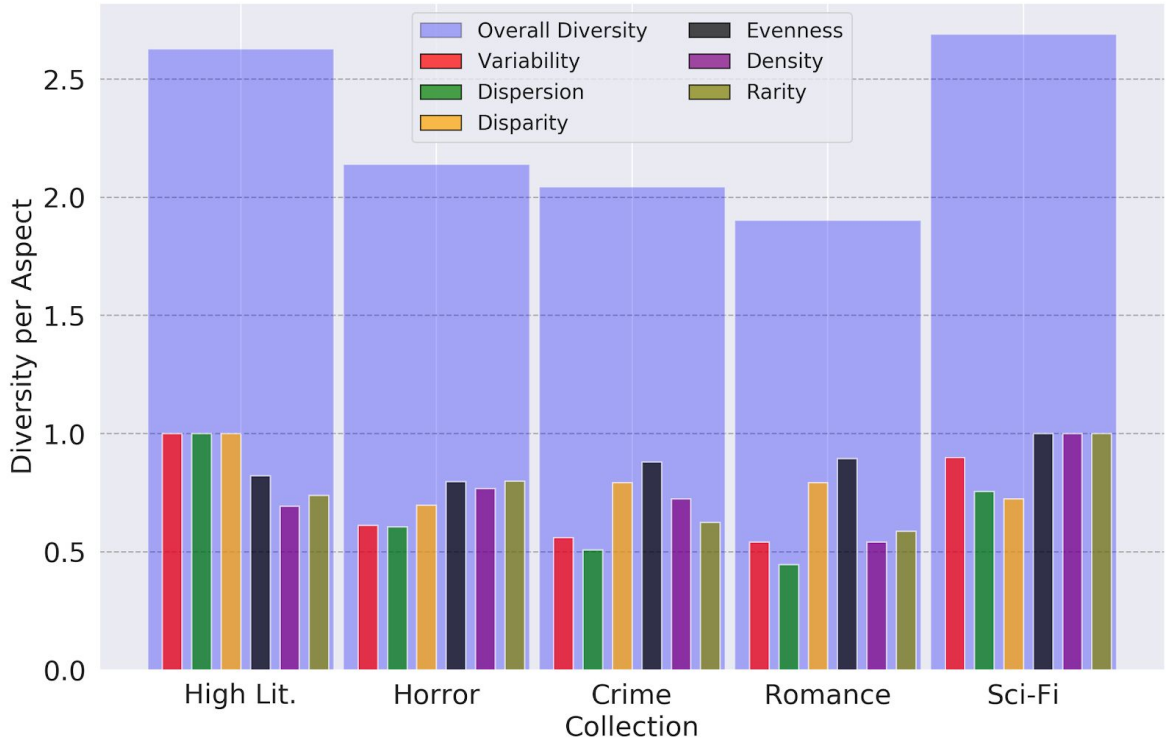

Fig 9: Diversity per aspect. For each dimension, the maximum value has been set to 1 and all others are proportional. Overall diversity is the sum of all dimensions (divided by 2).

Future work

In addition to the implementation of the yet missing measures, further testing of the influence of parameters like window size on the ranking and explanations for ranking differences are important. We also need to explore how to incorporate specific genre vocabulary into these measures.

References

Da, N. Z. (2019): The computational case against computational literary studies. Critical Inquiry, 45(3), p. 601–639.

Falk, I., Bernhard, D., Gerard. C. (2014): From Non Word to New Word: Automatically Identifying Neologisms in French Newspapers. In Proceedings of LREC 2014.

Jarvis, S. (2013a): Capturing the Diversity in Lexical Diversity. In: Language Learning 63 (1), p. 87–106.

Jarvis, S. (2013b): Defining and Measuring Lexical Diversity. In: Jarvis, Scott / Daller, Michael (Hrsg.): Vocabulary Knowledge. Human Ratings and Automated Measures. Amsterdam: John Benjamins. (= Studies in Bilingualism 47)

Klosa, A., Lungen, H. (2018): New German Words: Detection and Description. In Proceedings of the XVIII EURALEX, p. 559–569. Ljubljani.

Koschorke, A. (2016): Komplexität und Einfachheit. p. 1–10. Stuttgart.

Latcombe, G. and McGeosh, A. M. (2017): Capturing the Contribution of Rare and Common Species to Turnover: A Multi-Site Version of Generalised Dissimilarity Modelling. In Technological Advances at the Interface between Ecology and Statistics, Vol. 8/4. p. 393–526

Ney, H., Essen, U. Kneser, R. (1994): On structuring probabilistic dependences in stochastic language modelling In: Computer Speech & Language, Volume 8, Issue 1, p. 1-38.

Pielou, E.C. (1966): "The measurement of diversity in different types of biological collections". Journal of theoretical biology. 13: p. 131–144. doi:10.1016/0022-5193(66)90013-0

Schäfer, R. (2015): Processing and Querying Large Web Corpora with the COW14 Architecture. In: Proceedings of Challenges in the Management of Large Corpora (CMLC-3) (IDS publication server), p. 28–34.

Schäfer, R.& Bildhauer, F. (2012): Building Large Corpora from the Web Using a New Efficient Tool Chain. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), p. 486–493.