Piotrowski, M. and Fafinski, M. (2020). Nothing New Under the Sun? Computational Humanities and the Methodology of History. *Proceedings of the Workshop on Computational Humanities Research (CHR 2020)*, vol. 2723. Amsterdam: CEUR Workshop Proceedings, pp. 171–181 doi: urn:nbn:de:0074-2723-3. http://ceur-ws.org/Vol-2723/short16.pdf (accessed 5 October 2021).

Potter, W. J. (2010). The State of Media Literacy. *Journal of Broadcasting & Electronic Media*, **54**(4). Routledge: 675–696 doi: 10.1080/08838151.2011.521462. https://doi.org/10.1080/08838151.2011.521462 (accessed 5 October 2021).

Ramsay, S. (2013a). On building. In Vanhoutte, E., Nyhan, J. and Terras, M. (eds), *Defining Digital Humanities: A Reader*. Surrey: Ashgate, pp. 243–46 https://web.archive.org/web/20131205051042/http://stephenramsay.us/text/2011/01/11/on-building/ (accessed 6 December 2021).

Ramsay, S. (2013b). Who's In and Who's Out. In Vanhoutte, E., Nyhan, J. and Terras, M. (eds), *Defining Digital Humanities: A Reader*. Surrey: Ashgate, pp. 239–42 https://web.archive.org/web/20121015012254/http: //stephenramsay.us/text/2011/01/08/whos-in-and-whos-out.html (accessed 6 December 2021).

Spante, M., Hashemi, S. S., Lundin, M. and Algers, A. (2018). Digital competence and digital literacy in higher education research: Systematic review of concept use. (Ed.) Wang, S. *Cogent Education*, **5**(1). Cogent OA: 1519143 doi: 10.1080/2331186X.2018.1519143. https://doi.org/10.1080/2331186X.2018.1519143 (accessed 5 October 2021).

Tafazoli, D., Parra, M. E. G. and Abril, C. A. H. (2017). Computer literacy: Sine qua non for digital age of language learning & teaching. *Theory and Practice in Language Studies*, **7**(9). Academy Publication Co., Ltd.: 716 doi: 10.17507/tpls.0709.02. https://www.researchgate.net/profile/Dara-Tafazoli/project/My-PhD-Thesis-A-cross-cultural-study-on-the-relationship-between-CALL-literacy-and-the-attitudes-of-Spanish-and-Iranian-English-language-students-and-teachers-towards-CALL/attachment/59bc435a4cde26fd91fbe78c/AS:538963140988929@1505510234363/download/1248-4859-1-PB.pdf?context=ProjectUpdatesLog (accessed 5 October 2021).

Tannenbaum, R. S. (1987). How Should We Teach Computing to Humanists?. *Computers and the Humanities*, **21**(4). Springer: 217–25 http://www.jstor.org/stable/30207392 (accessed 6 December 2021).

Timans, R., Wouters, P. and Heilbron, J. (2019). Mixed methods research: what it is and what it could be. *Theory and Society*, **48**(2): 193–216 doi: 10.1007/s11186-019-09345-5. https://doi.org/10.1007/s11186-019-09345-5 (accessed 6 December 2021).

Van Zundert, J. J., Antonijević, S. and Andrews, T. L. (2020). 'BlackBoxes' andTrue Colour — A Rhetoric of Scholarly Code. In Edmond, J. (ed), *Digital Technology and the Practices of Humanities Research*. Cambridge, UK: Open Book Publishers, pp. 123–62 https://www.openbookpublishers.com/product/1108 (accessed 17 February 2020).

Vee, A. (2013). Understanding Computer Programming as a Literacy. *LiCS*, **1**(2): 42–64 https://licsjournal.org/index.php/LiCS/article/view/794 (accessed 30 September 2021).

Vee, A. (2017). *Coding Literacy: How Computer Programming Is Changing Writing*. (Software Studies). Cambridge: The MIT Press.

Webster, S. W. and Webster, L. S. (1985). Computer Literacy or Competency?. *Teacher Education Quarterly*, **12**(2). Caddo Gap Press: 1–7 http://www.jstor.org/stable/23474573 (accessed 6 December 2021).

Exploring Lexical Diversities

Blombach, Andreas
andreas.blombach@fau.de
University of Erlangen-Nürnberg, Germany

Evert, Stephanie
stephanie.evert@fau.de
University of Erlangen-Nürnberg, Germany

Jannidis, Fotis
fotis.jannidis@uni-wuerzburg.de
University of Würzburg, Germany

Pielström, Steffen
pielstroem@biozentrum.uni-wuerzburg.de
University of Würzburg, Germany

Konle, Leonard
leonard.konle@uni-wuerzburg.de
University of Würzburg, Germany

Proisl, Thomas
thomas.proisl@fau.de
University of Erlangen-Nürnberg, Germany

The validation corpora (Weiß & Meurers 2018) 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 (161 works, approx. 17 mio. tokens) with “dime novels” (1167 works in six different genres, approx. 40 mio. tokens), both under copyright. Dime novels are a type of fiction mass-produced in long-lasting series and sold in kiosks rather than book stores.

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 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).

TTR depends systematically on sample size. Among the solutions proposed for this problem, standardized TTRs (STTR) calculated from fixed-length text chunks provide a practical and intuitive solution (Fig. 1).

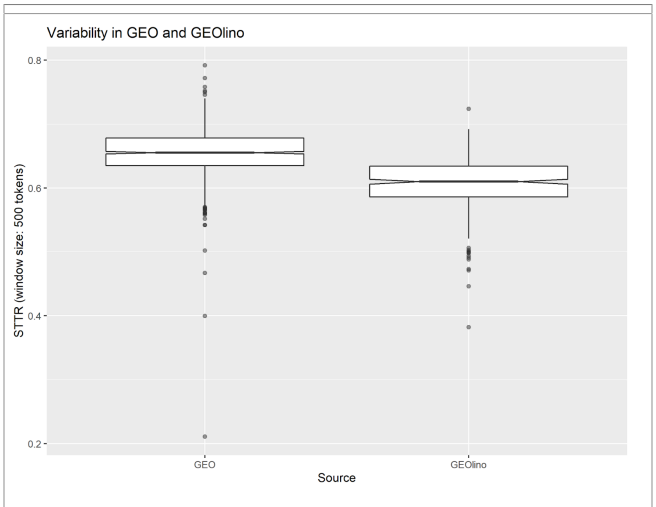

Figure 1: STTR in GEO and GEOlino

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 very common words. We use a simple approach to model rarity. For each text, we compute the proportion of content words not included in the 5,000 most frequent content words from a large web corpus that covers many different registers, the DECOW16BX (Fig. 2, Schäfer and Bildhauer 2012, Schäfer 2015).

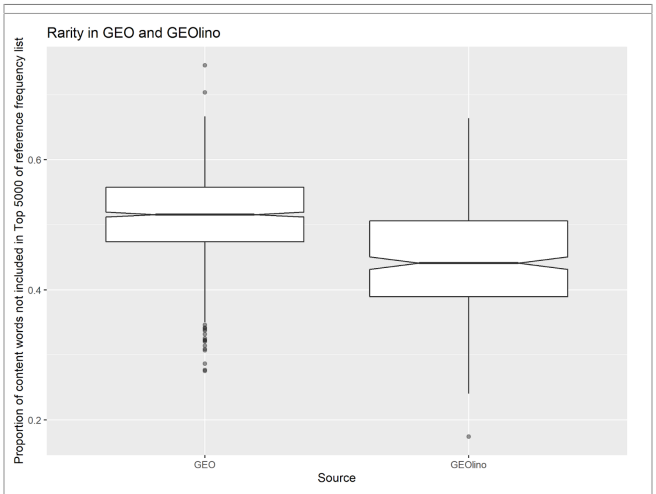

Figure 2: *Rarity* in GEO and GEOlino

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 produces an impression of redundancy. To measure this effect, we again use a window-based approach (Fig. 3). Inside a window, we calculate a dispersion score based on the Gini coefficient (Gini 1912) for each type and use the arithmetic mean of this score over all types with a frequency greater than one as dispersion measure for the whole text (see Blombach et al. in preparation for a detailed description).

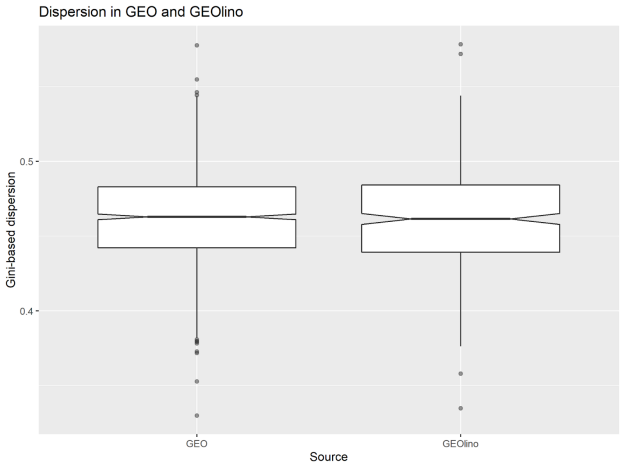

Figure 3: Dispersion in GEO and GEOlino

Disparity

Lexical disparity follows the intuition that repetition also shows in the occurrence of similar words on a semantic level. To measure global disparity, a document is segmented and a vector is then generated for each segment by averaging over the vectors of the content words. The disparity of a segment is then calculated from the pairwise euclidean distance of all its segments. The document's disparity is the mean over all its segment disparities (Fig. 4).

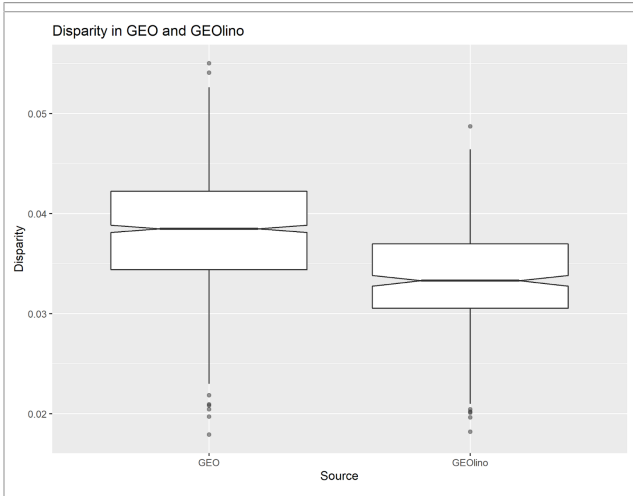

Figure 4: Disparity in GEO and GEOlino

Density

A text containing a higher proportion of content words can be considered denser and therefore more complex (Fig. 5).

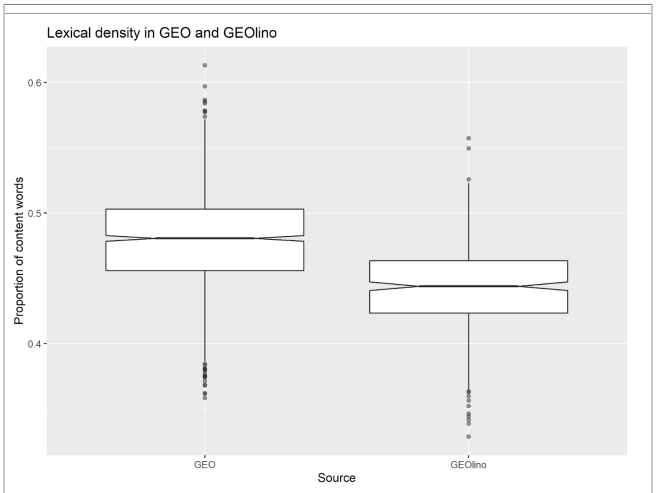

Figure 5: Density in GEO and GEOlino

Tools

Most of the measures suggested here (variability, rarity, dispersion, and density) are implemented in our textcomplexity toolbox that contains additional complexity measures as well.

We have also created an interactive “ Shiny” app which allows users to visually explore our data, including correlations between different measures and the influence of parameters such as window size, case sensitivity and the inclusion or exclusion of punctuation.

Application to Literature

Fig. 6 shows the measures of lexical complexity applied to six genres of dime novels and a set of highbrow novels. Counter to our expectations, science fiction and fantasy equal or even surpass the highbrow novels in some respects (disparity, density, dispersion and rarity). We assume that we have different forms of lexical complexity at work here: In science fiction and fantasy, a noun-heavy prose is depicting new worlds with new words. In high literature on the other hand, high variability shows the influence of a stylistic ideal which aims to avoid repetition and show elegance. There might be a difference in the scope which authors control for complexity, for example variability. We found less repetition in small windows in genre texts, whereas variability in highbrow literature increases with window size.

Fig. 7 shows that genre similarities can be perceived immediately using this kind of representation. A multi-dimensional model of lexical complexity allows a clearer understanding of genre differences.

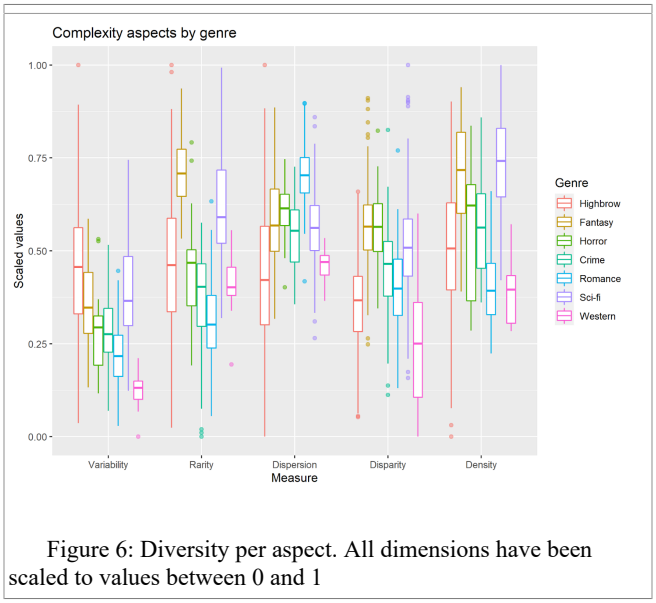

Figure 6: Diversity per aspect. All dimensions have been scaled to values between 0 and 1

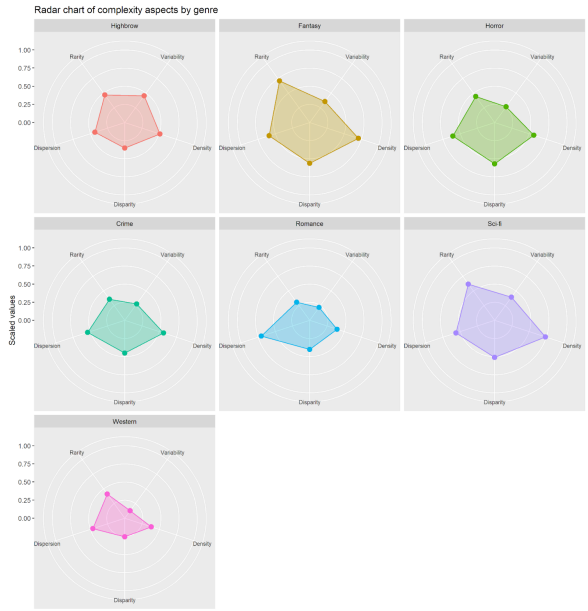

Figure 7.: Radarplots, highlighting the similarities between genres

Bibliography

Blombach, A., Evert, S., Jannidis, F., Konle, L., Pielström, S. and Proisl, T. (in preparation): Lexical Complexity in Texts. A Multidimensional Model.

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

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

Gini, C. (1912): *Variabilità e Mutuabilità. Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche*. C. Cuppini, Bologna.

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 (Eds.): *Vocabulary Knowledge. Human Ratings and Automated Measures*. Amsterdam: John Benjamins. (= Studies in Bilingualism 47)

Klosa, A. and 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.

Ney, H., Essen, U. and 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. In: *Journal of theoretical biology*. 13: p. 131–144. doi:10.1016/0022-5193(66)90013-0

Pielström, S., Hodošček, B., Calvo Tello, J., Henny-Krahmer, U., Jannidis, F., Schöch, C., Du, K., Uesaka, A. and Tabata, T. (in preparation): Measuring Lexical Diversity of Literary Texts.

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. and 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.

Weiß, Z. and Meurers, D. (2018): Modeling the readability of German targeting adults and children: An empirically broad analysis and its cross-corpus validation. In: *Proceedings of the 27th International Conference on Computational Linguistics*, p. 303–317, Santa Fe, New Mexico, USA.