A new approach to measuring lexical diversity in historical corpora.

Lauren Fonteyn[1,2], Folgert Karsdorp[2] & Enrique Manjavacas[1]
[1]University of Leiden, l.fonteyn@hum.leidenuniv.nl [2] KNAW Meertens Institute Amsterdam

Keywords: Lexical Diversity, Productivity, Word Embeddings, Corpus Linguistics, Computational Linguistics

The question whether and how we can measure lexical diversity has long been a pertinent one in Linguistics and related disciplines. Attempts have been made to estimate the vocabulary size of (average speakers of) a particular language (at different ages) (e.g. Ellegård 1960, Brysbaert et al. 2016, Segbers & Schroeder 2017), and many studies in (Diachronic) Construction Grammar are concerned with estimating the number of unique lexical items that may occur in particular morphosyntactic structures for different individuals or across time (e.g. Schmid & Mantlik 2015; Perek 2018). To address these questions, researchers often resort to corpus research, using quantitative measures that rely on type and token frequency and/or hapax legomena, such as (variations on) Mean Word Frequency (MWF) and Type-Token Ratio (TTR) (see Tweedie & Baayen 1998), and realized/potential/expanding productivity (Baayen 2009).

However, in many corpora, unique character strings cannot be equated to unique words. This may be due to spelling variation or OCR errors, which are very common in historical corpora (e.g. the Modern English character <ſ> is often mistaken for <f> or <l>, which means *strength* <ſtrength> can also be represented by <frength> and <lrength>). Because neither OCR errors nor non-standard spelling variation are entirely systematic, reducing such variation through corpus pre-processing can be challenging. As a solution, we propose an approach originally developed to estimate ecological diversity (Chao et al. 2019) called the attribute diversity framework, which distinguishes categorical diversity from functional diversity. We define 'categorical diversity' as the number of unique words in a text, and 'functional diversity' as a measure that also takes into account their distributional similarity. Operationalizing this similarity by means of word embeddings generated with the historically pre-trained language model MacBERTh (Manjavacas & Fonteyn 2022), we demonstrate that:

(i) Functional diversity estimates are affected to a much lesser extent by spelling inconsistencies and OCR errors than categorical diversity.

(ii) Given two sets of unique word types, set A{*dog, bird, rabbit*} and set B{*progesterone, remember, blue*}, the approach also captures the higher functional-semantic diversity of set B.

As a concrete case study to demonstrate the theoretical and practical advantages of discussing 'vocabulary richness' in terms of attribute diversity, we use the diachronic ARCHER corpus (version 3.2) and discuss diachronic changes in and differences between texts from different genres and by different authors in terms of categorical as well as functional diversity.

References

Baayen, R. Harald. 2009. Corpus linguistics in morphology: Morphological productivity. in: A. Lüdeling, M. Kytö (eds.), *Corpus Linguistics: An International Handbook*, 899–919. Berlin: De Gruyter.

Brysbaert, Marc, Michaël Stevens, Pawel Mandera & Emmanuel Keuleers. 2016. How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in Psychology* 7.

Chao, Anne, Chun-Huo Chiu, Sébastien Villéger, I-Fang Sun, Simon Thorn, Yi-Ching Lin, Jyh-Min Chiang & William B. Sherwin. 2019. An attribute-diversity approach to functional diversity, functional beta diversity, and related (dis)similarity measures. *Ecological Monographs* 89.

Ellegård, Alvar. 1960. Estimating vocabulary size. *WORD* 16: 219–244.

Manjavacas, Enrique & Lauren Fonteyn. 2022. Adapting vs. Pre-training Language Models for Historical Languages. *Journal of Data Mining & Digital Humanities*. jdmdh:9152.

Schmid, Hans-Jörg & Annette Mantlik. 2015. Entrenchment in historical corpora? Reconstructing dead authors' minds from their usage profiles. *Anglia* 133: 583–623.

Segbers, Jutta & Sasha Schroeder. 2017. How many words do children know? A corpus-based estimation of children's total vocabulary size. *Language Testing* 34: 297–320.

Perek, Florent. 2018. Recent change in the productivity and schematicity of the *way* -construction: A distributional semantic analysis, *Corpus Linguistics and Linguistic Theory* 14: 65–97.

Tweedie, Fiona J. & R. Harald Baayen. 1998. How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities* 32: 323–352.