

Standardization and orthographic variation in Late Modern Dutch witness depositions

Sara Budts, Yoshi Malaise & Rik Vosters | Vrije Universiteit Brussel, Belgium

In this contribution, we will present a quantitative analysis of orthographic variation in the *Witnesses* corpus, a collection of Flemish witness depositions from the 18th and 19th centuries, interpreting the findings from the perspective of linguistic standardization (Rutten & Vosters, 2021).

Most studies quantifying spelling standardization as a diachronic process depart from corpus normalization and lemmatization, measuring differences between observed spelling forms and modern standard language equivalents (e.g. Tyrkkö 2020). This approach, however, presupposes a teleological model of standardization, where historical variation is reduced over time only to gradually move closer to an invariable modern standard, as part of an inevitable linear trajectory from ‘dialect’ to ‘standard’ — criticized by Watts (2011) as the ‘funnel view’ of standardization.

In our contribution, we will present a novel heuristic approach to quantify orthographic variation, avoiding the pitfall of teleological bias. We generate word embeddings using a pre-trained historical language model, fine-tuned on the *Witnesses* material, and combine these with Levenshtein edit distances between a particular token and its semantic nearest neighbors in embedding space, to identify orthographic variants of the same lexical item in a set of texts. We then measure the amount of orthographic variation across texts, investigating diachronic trends as well as differences between text types, scribes, and across other sociolinguistic variables.

The results of this bottom-up approach will then be interpreted in light of the standardization history of Dutch, with a particular focus on the Southern Low Countries. Although the proposed method provides us with relatively coarse-grained quantitative metrics of orthographic variation in historical texts, we will show how our findings align with earlier work on spelling change, focusing on features undergoing standardization-driven shifts in the same period.

Rutten, G., & Vosters, R. (2021). Language standardization ‘from above.’ In W. Ayres-Bennett & J. Bellamy (Eds.), *The Cambridge Handbook of Language Standardization* (pp. 65–92). Cambridge University Press.

Tyrkkö, J. (2020). Early modern medicine in manuscript and print: A triangulation approach to analysing spelling standardisation. *International Journal of English Studies*, 20(2), 67–93.

Watts, R. (2011). *Language myths and the history of English*. Oxford University Press.