An information-theoretic approach to morphological and syntactic complexity in Dutch, English and German

Julie Nijs, Freek Van de Velde, and Huybert Cuyckens

Larger languages in high-contact communities are morphologically less complex and rely more on lexical strategies and word order than smaller languages in close-knit communities (Lupyan & Dale 2010). This study focuses on the West-Germanic languages Dutch, English and German, which are known to have been exposed to different degrees of internal (dialect) contact and external contact (O'Neil 1978; Weerman 2006). Specifically, English has been more exposed to contact than Dutch, which in turn has been more exposed than German. To assess whether degree of contact correlates with morphological as well as syntactic complexity in these languages, we measure morphological and syntactic complexity by the mathematical notion of 'Kolmogorov complexity' (Kolmogorov 1968), an information-theoretic approach which defines a string's complexity in relation to its information content.

The Dutch, English and German texts making up our dataset were taken from the Book of Genesis and the Gospel of Matthew, as they occur in the multilingual parallel EDGeS Diachronic Bible Corpus (Bouma, Coussé, Dijkstra & van der Sijs 2020). A total of 47 texts from different time periods between the 14th and 19th century have been analyzed: 21 for Dutch, 18 for English and 8 for German.

Following Juola (2008) and Ehret (2017), morphological complexity can be calculated after randomly deleting 10% of a text's orthographic transcribed characters and compressing the file with gzip. The random deletion leads to morphological distortion, in that the number of unique tokens increases, which makes compressibility worse. Texts characterized by a high surface token diversity (as a result of affixal complexity, root-internal alternation or other morphological operations) will be comparatively less affected by distortion, because they already contain a higher amount of unique tokens before distortion. In terms of Kolmogorov complexity, these are the texts that are morphologically more complex. Syntactic complexity can be calculated in the same way, but instead of characters, words are deleted. This leads to a distortion of the word order rules, a higher number of unique lexical n-grams and thus worse compressibility. Texts with strict word order have more structural surface redundancies and will therefore be more affected by distortion, while languages with free word order will be less affected due to their lower number of redundancies. This means that in terms of Kolmogorov complexity rigid word order is considered as more complex.

The morphological complexity ratio is calculated as $-\frac{mc}{c}$, where mc is the compressed file size in bytes after morphological distortion, and c is the compressed file size in bytes before distortion. The syntactic sc complexity ratio or the word order rigidity ratio is calculated as , where sc is the compressed file size c in bytes after syntactic distortion, and c is the compressed file size in bytes before distortion. For each text the mean morphological and syntactic complexity was calculated over 1000 iterations, to take the aleatoric effect of the randomization into account.

We have found a significant interaction effect between year and language for the morphological complexity ratio. Morphological simplification happens faster in English compared to Dutch, as expected, but German seems to be more on the side of English, counter to what we expect. Syntactic complexity, then, shows the mirror image. We can thus observe a negative correlation between the morphological and syntactic complexity ratio (Figure 1). The three languages each take up their own space in the graph. Dutch is morphologically the most complex, but syntactically less complex; English is syntactically the most complex, but morphologically less complex; German lies in-between.

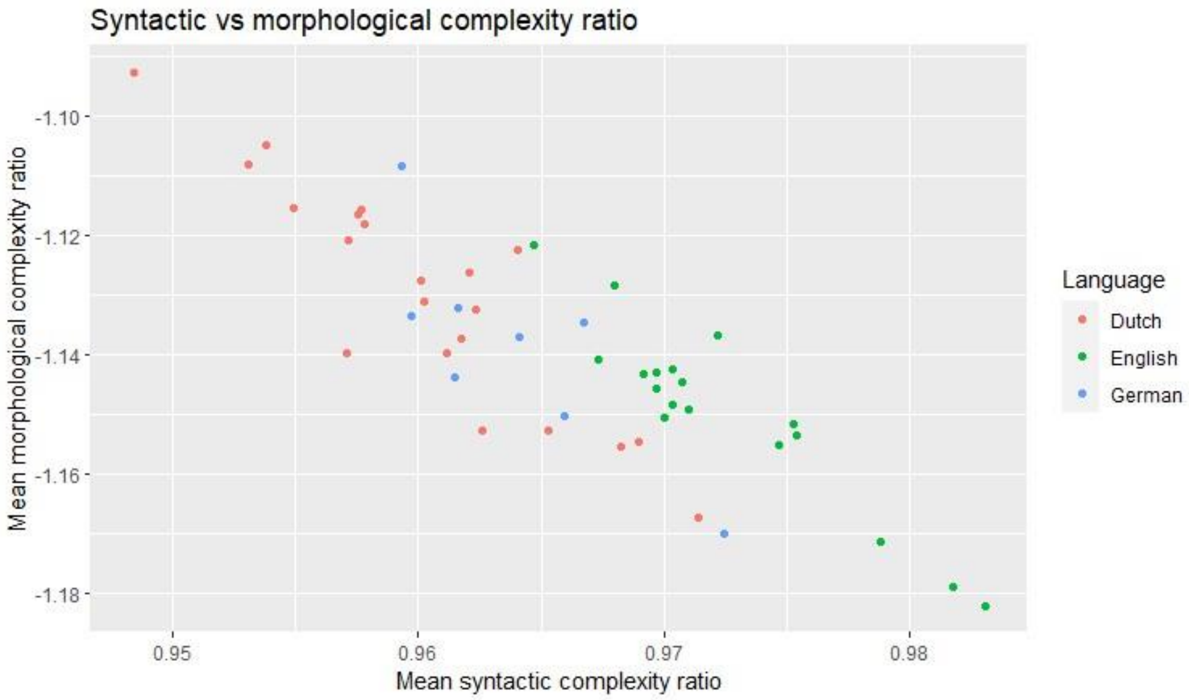

Figure 1: Syntactic vs morphological complexity ratio

References

Bouma, G., E. Coussé, T. Dijkstra & N. van der Sijs. 2020. The EDGeS diachronic bible corpus. In *Proceedings of the 12th international conference on language resources and evaluation*, 5232-5239. Paris: ELDA.

Ehret, K. 2017. *An information-theoretic approach to language complexity: variation in naturalistic corpora*. Freiburg: Albert-Ludwig-Universität Freiburg dissertation.

Juola, P. 2008. Assessing linguistic complexity. In M. Miestamo, K. Sinnemäki & F. Karlsson (eds.), *Language complexity: typology, contact, change*, 89-108. Amsterdam: John Benjamins.

Kolmogorov, A. Ni. 1968. Three approaches to the quantitative definition of information. *International Journal of Computer Mathematics 2*(1-4). 157-168. doi:10.1080/00207166808803030.

Lupyan, G. & R. Dale. 2010. Language structure is partly determined by social structure. *PLoS One 5*(1).

O'Neil, W. 1978. The evolution of the Germanic inflectional systems: a study in the causes of lan-guage change. *Orbis* 27: 248-286.

Weerman, F. 2006. It’s the economy, stupid: Een vergelijkende blik op *men* en *man*. In: M. Hüning, et al. (eds.), *Nederlands tussen Duits en Engels*. SNL. 19-47.