Comparing the internal complexity of words: A pilot study on parallel texts in seven languages

Keywords: morphological segmentation; internal structure; morph; morpheme; complexity

The comparison of languages at the level of words and their internal structure is an intrinsic part of contrastive linguistics and language typology and has led, among other things, to the well-known morphological typology in terms of agglutination, fusion, isolation, etc. (Skalička, 1935; Sgall, 1986). Unlike syntactic research, where typological claims are refined and advanced on the basis of multilingual syntactically annotated corpora (in particular, Universal Dependencies, de Marneffe et al. 2021; cf., among others, Choi et al. 2022 or Levshina et al. 2023), theoretical insights related to the internal structure of words have not yet been explored on a broad data base. One reason is the inaccessibility of the words' structure in machine tractable data.

Analysis of the internal structure of words in terms of their segmentation into morphemes as the smallest meaning-bearing units (so-called morphological segmentation) is available in printed dictionaries for some of the languages involved in the present experiment (Slavíčková 1975; Ološtiak et al. 2015; Kuznetsova and Efremova 1986; Tikhonov 1996). Information on the words' internal is also available in some dedicated electronic resources. However, their utility for language comparisons is limited because they either provide reliable but mutually incompatible analyses for individual languages (e.g. CELEX, Baayen et al. 1995, or MorphoLex, Sánchez-Gutiérrez et al. 2018 and Mailhot et al. 2020) or, if they attempt multilingual coverage, the data for individual languages are of poor quality and inconsistent across languages (UniMorph, McCarthy et al. 2020; or UniSegments, Žabokrtský et al. 2022).

We report a pilot study in which we compare the complexity of words across seven typologically diverse languages (Czech, English, Finnish, French, German, Russian, and Slovak). The study is carried out on the texts of *Universal Declaration of Human Rights*, which, in each language, consists of a preamble and 30 articles sized from one sentence to several short paragraphs. The texts are segmented into morphemes and these are classified into roots and affixes, which are further divided into inflectional and derivational. Both morphological segmentation and morpheme classification are performed by neural tools using customized CNN-LSTM-CRF architecture (Ma and Hovy, 2016). The tools were trained on data extracted from the above mentioned resources.

Despite the relatively small size of the texts and the automatic processing, which is being further improved, the analysis provides interesting insights. In all languages, single-morpeme words are the most numerous, followed by words with two morphemes, etc. Figure 1 documents that the tendency for word length to be inversely related to the word's relative frequency (discussed as Zipf's law) is more stable when word length is calculated in terms of morphemes than in terms of characters. The data also show differences in the number and diversity of roots and derivational morphemes. Based on this metric, languages cluster according to genetic relationships. The data thus exhibit the potential for comparing languages as for their preferences for different word-formation processes.

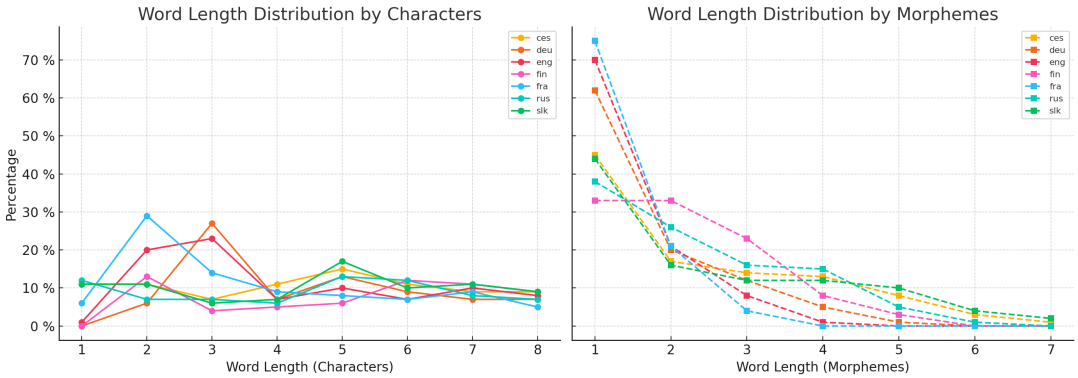

Figure 1: Number of tokens with given number of morphemes/characters

References

Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1995). The CELEX lexical database (CD-ROM). Catalogue No. LDC96L14.

Choi, H.-S., Guillaume, B., and Fort, K. (2022). Corpus-based language universals analysis using universal dependencies. In *roceedings of the Second Workshop on Quantitative Syntax (Quasy, SyntaxFest 2021)*, pages 33–44.

de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Kuznetsova, A. I. and Efremova, T. F. (1986). *Slovar' morfem russkogo jazyka [Dictionary of morphemes of the Russian language]*. Russkij jazyk, Moscow.

Levshina, N. et al. (2023). Why we need a gradient approach to word order. *Linguistics*, 61(4):825–883.

Ma, X. and Hovy, E. (2016). End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In Erk, K. and Smith, N. A., editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin. ACL.

Mailhot, H. et al. (2020). MorphoLex-FR: A derivational morphological database for 38,840 French words. *Behavior Research Methods*, 52(3):1008–1025.

McCarthy, A. D. et al. (2020). Unimorph 3.0: Universal morphology. In *Proceedings of the 12th Language Resources and Evaluation Conference*. ELRA.

Ološtiak, M., Genči, J., Rešovská, S., and univerzita. Filozofická fakulta, P. (2015). *Retrográdny morfematický slovník slovenčiny*. Acta Facultatis Philosophicae Universitatis Prešoviensis: Slovník. Filozofická fakulta Prešovskej univerzity v Prešove.

Sgall, P. (1986). Classical typology and modern linguistics. *Folia linguistica*, 20:15–28.

Skalička, V. (1935). *Zur ungarischen Grammatik*. Nakladatelství University Karlovy, Praha.

Slavíčková, E. (1975). *Retrográdní morfematický slovník češtiny: s připojenými inventárními slovníky českých morfémů kořenových, prefixálních a sufixálních*.

Sánchez-Gutiérrez, C. H. et al. (2018). MorphoLex: A derivational morphological database for 70,000 English words. *Behavior Research Methods*, 50(4):1568–1580.

Tikhonov, A. N. (1996). *Morfemno-orfografičeskij slovar' russkogo jazyka. Russkaja morfemika (Morphemic-spelling dictionary of the Russian language. Russian morphemics)*. Shkola-Press, Moscow, Russia.

Žabokrtský, Z. et al. (2022). Towards universal segmentations: UniSegments 1.0. In Calzolari, N. et al., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1137–1149, Marseille. ELRA.