Geda Paulsen/Ene Vainik/Maria Tuulik/Ahti Lohk

THE MORPHOSYNTACTIC PROFILE OF PROTOTYPICAL ADJECTIVES IN ESTONIAN

Keywords Lexical categories; morphosyntax; lexicography; language technology; Estonian

The current direction in Estonian lexicography is a unification of lexical resources (dictionaries and term bases) into a central superdictionary, the EKI Combined Dictionary (CombiDic), supported by the dictionary writing system Ekilex. The lexicographic work is moving towards a higher degree of automation and processing of corpora (Koppel et al. 2019; Tavast et al. 2020) The lexical database of Ekilex includes automatically generated lists over dictionary entry candidates, requiring assessment of their degree of lexicalisation. An urgent lexicographic issue is providing the underspecified Ekilex entries with PoS tags and assessing the candidates for their potential status as a lexical entry. Today, 72% of the total number of the public CombiDic headwords miss the PoS tag.

In this study, we focus on one of the most ambiguous parts of speech posing categorisation problems for the lexicographers, the adjective (Paulsen et al. 2019, p. 327). To clarify this issue, we aim to develop a multi-parameter solution for determining the relative adjectiveness of a word or a word form, e. g., the adjectivizing participles or nominals (for the border areas of adjectives with other lexical classes in Estonian, see Vainik/Paulsen/Lohk 2020).

In our vision, the properties of a concrete word can be compared with the profile of a typical adjective, using the profile as a similarity measure. We have established and tested the characteristic attributes of the adjective in a previous study (Tuulik et al. in press), using six morphosyntactic parameters detectable in the corpus. The result of the experiment was that the tested parameters were, to different degrees, able to differentiate adjectival morphosyntactic behaviour.

To establish the exact boundaries of the profile of prototypical adjectives, the parameters should be tested on a larger sample of adjectives that represent the best examples of its category. Our aim in the present study is to find the representative profile of the prototypical adjective based on a larger sample of predefined adjectives and setting the threshold value for classifying the questionable word form as an adjective. The normal distribution of the parameter values will be used to distinguish adjectives from other words and used as a comparison to the corresponding values of the unclear cases.

The basis for the analysis is the largest corpus of contemporary Estonian, the Estonian National Corpus 2019 with 1.5 billion words. The corpus is lemmatised, tagged and disambiguated with the EstNLTKv.1.6 toolkit (Laur et al. 2020). The selection of test words contains 100 words extracted by random sampling from the 554 most central Estonian adjectives included in the Basic Estonian Dictionary (Kallas et al. 2014). The Euclidean distance analysis calculated from the profile of the prototypical adjective is the measure of a word form's similarity vs. difference according to its behaviour in the corpus.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Book of Abstracts of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

References

Laur, S./Orasmaa, S./Särg, D./Tammo, P. (2020): EstNLTK 1.6: Remastered Estonian NLP Pipeline. In: Proceedings of the 12th Language Resources and Evaluation Conference. Marseille, pp. 7152–7160.

Ekilex (2022): https://ekilex.eki.ee/ (last access: 20-03-2022).

Estonian National Corpus 2019 = Koppel, K./Kallas, J. (2020): Eesti keele ühendkorpus 2019. https://doi.org/10.15155/3-00-0000-0000-0000-08565L.

CombiDic = Hein, I./Kallas, J./Kiisla, O./Koppel, K./Langemets, M./Leemets T./Melts, M./Mäearu, S./Paet, T./Päll, P./Raadik, M./Tiits, M./Tsepelina, K./Tuulik, M./Uibo, U./Valdre, T./Viks, Ü./Voll, P. (2020): The EKI Combined Dictionary. Institute of the Estonian Language. https://sonaveeb.ee (last access: 20-03-2022).

Kallas, J./Tiits, M./Tuulik, M./Koppel, K./Jürviste, M. (2014): Eesti keele põhisõnavara sõnastik [The Basic Estonian Dictionary]. Tallinn.

Koppel, K./Tavast, A./Langemets, M./Kallas, J. (2019): Aggregating dictionaries into the language portal Sõnaveeb: issues with and without a solution. In: Kosem, I./Zingano Kuhn, T./Correia, M./Ferreria, J. P./Jansen, M./Pereira, I./Kallas, J./Jakubíček, M./Krek, S./Tiberius, C. (eds.): Proceedings of the eLex 2019 Conference. 1–3 October 2019, Sintra, Portugal. Brno, pp. 434–452.

Paulsen, G./Vainik, E./Tuulik, M./Lohk, A. (2019): The lexicographer's voice: word classes in the digital era. In: Kosem, I./Zingano Kuhn, T./Correia, M./Ferreria, J. P./Jansen, M./Pereira, I./Kallas, J./Jakubíček, M./Krek, S./Tiberius, C. (eds.): Proceedings of the eLex 2019 Conference. 1–3 October 2019, Sintra, Portugal. Brno, pp. 434–452.

Vainik, E./Paulsen, G./Lohk, A. (2020): A typology of lexical ambiforms in Estonian. In: Gavriilidou, Z./Mitsiako, M./Fliatouras, A. (eds.): Lexicography for Inclusion. Proceedings of the 19th EURALEX Congress, 7-9 September 2021, Alexandroupolis. Volume 1. Alexandroupolis, pp. 119–130.

Tuulik, M./Vainik, E./Lohk, A./Paulsen, G. (in press): Kuidas ära tunda adjektiivi? Korpuskäitumise mustrite analüüs [How to recognize adjectives? An analysis of corpus patterns]. The Estonian Papers in Applied Linguistics.

Tavast A./Koppel, K./Langemets, M./Kallas, J. (2020): Towards the Superdictionary: Layers, Tools and Unidirectional Meaning Relations. In: Gavriilidou, Z./Mitsiako, M./Fliatouras, A. (eds.): Proceedings of XIX EURALEX Congress: Lexicography for Inclusion, Volume 1., Greece, pp. 215–223.

Contact information

Geda Paulsen
Institute for the Estonian Language/Uppsala University
geda.paulsen@eki.ee

Ene Vainik
Institute for the Estonian Language
ene.vainik@eki.ee

Maria Tuulik
Institute for the Estonian Language
maria.tuulik@eki.ee

Ahti Lohk
Institute for the Estonian Language
ahti.lohk@eki.ee

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Book of Abstracts of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Acknowledgements

This work is supported by Estonian Research Council grant PSG227.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Book of Abstracts of the XX EURALEX International Congress. Mannheim: IDS-Verlag.