

Linguistic variation in written standard Finnish: a corpus-driven approach

Elisa Reunanen & Ilmari Ivaska

University of Turku

Different kinds of texts differ in their context of use, layout, and communicative purposes. This study investigates the variation of standard Finnish language across four genres (research articles, fiction, political programs, and news) by combining two computational methods. Linguistic analysis of a genre is grounded in the typical features of the language register used in that genre (Biber & Conrad 2019). The concepts of genre and register are employed as follows: genre refers to a category defined by criteria external to language, while register describes the variation in linguistic features within a text or its sections, in relation to the meanings they convey. The categorization of texts into genres and registers is based on shared and distinguishing features and their functional interpretations, with a feature considered typical of one genre only when compared to others. Hence, linguistic features that characterize a given genre in relation to other genres, can be expected to reflect the communicative goals of the text. Earlier research on English (Biber et al. 2021) and on some other languages (e.g. Biber 2014) has identified five register dimensions, i.e. discourse functions. They explain the vast majority of language variation. Each dimension is a continuum, and each dimension has linguistic and functional content: linguistic content means, for example, morphological and syntactic features, while functional content refers to the communicative goals of the respective text.

This study is corpus-driven, which means that the studied features are selected from the bottom up by using computational methods. The research data stem from five different corpora (Corpus of Academic Finnish, Finnish News Agency Archive, Corpus of Translated Finnish, InterCorp, POHTIVA), and they comprise research articles, fictional texts, political programs, and news texts. Key structure analysis (Ivaska 2015) was employed to identify linguistic features significant in distinguishing between the subsets of data – features, typical to the different types of texts. These features were then grouped into sets according to their patterns of co-occurrence in the data, using multi-dimensional analysis (Egbert & Staples 2019). By combining key structure analysis and multi-dimensional analysis it is possible to identify linguistic features characteristic of each genre.

Research questions are:

- 1) Which linguistic features are characteristic of register used in each genre?
- 2) Which communicative goals are reflected by clusters of linguistic features?

Based on linguistic variation, three discourse dimensions are identified, reflecting the communicative goals of each genre. Research articles are characterized by the accumulation of modifiers at the beginning and end of sentences. In multi-dimensional analysis, the strongest positive association of research articles is at the dimension ‘abstract vs. non-abstract style’ and the strongest negative association is at the dimension ‘interactivity vs. accuracy’. Fiction is characterized by the versatile inflection of finite verb. In multi-dimensional analysis, the strongest positive association of fiction is at the dimension ‘interactivity vs. accuracy’ and the strongest negative association is at the dimension ‘influence vs. narrative’. Political programs are characterized by the use of noun modifiers and the abundance of both coordinating and subordinating clauses. In multi-dimensional analysis, the strongest association of political programs is at the dimension ‘influence vs. narrative’. News texts are characterized by the use of preterit, and the strongest association in multi-dimensional analysis is negative association at the dimensions ‘influence vs. narrative’ and ‘abstract vs. non-abstract style’.

All in all, the results indicate that there is considerable variation in standard Finnish language, and each studied genre employs a different linguistic register to achieve the communicative goals and to accommodate the particular circumstances of the respective genre.

References

- Biber, Douglas. 2014. Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Languages in Contrast*. John Benjamins 14(1). 7–34.
<https://doi.org/10.1075/lic.14.1.02bib>.
- Biber, Douglas & Susan Conrad. 2019. *Register, Genre, and Style*. 2nd edn. Cambridge University Press. <https://doi.org/10.1017/9781108686136>.
- Biber, Douglas, Bethany Gray, Shelley Staples & Jesse Egbert. 2021. *The Register-Functional Approach to Grammatical Complexity: Theoretical Foundation, Descriptive Research Findings, Application*. 1st edn. New York: Routledge.
<https://doi.org/10.4324/9781003087991>.
- Egbert, Jesse & Shelley Staples. 2019. Doing multi-dimensional analysis in SPSS, SAS, and R. In *Multi-Dimensional Analysis: Research Methods and Current Issues*, 99–114. Bloomsbury Academic. <https://doi.org/10.5040/9781350023857>.
- Ivaska, Ilmari. 2015. *Edistyneen oppijansuomen korpuspiirteitä korpusvetoisesti: Avainrakenneanalyysi*. Turku: University of Turku PhD Thesis.
<http://urn.fi/URN:ISBN:978-951-29-6204-4>.