Factor analysis of Russian register and linguistic variation

Freitag
10.03.2017
13:30 – 14:00
B4 1, 0.04

Roland Meyer
Humboldt-Universität zu Berlin
roland.meyer@hu-berlin.de

Luka Szucsich
Humboldt-Universität zu Berlin
luka.szucsich@hu-berlin.de

Slavic register phenomena have traditionally been captured by a fixed set of functional styles (FS). With modern linguistic corpora, registers may be derived bottom-up by a factor analysis of the (non-)occurrence of selected linguistic features across text types (Biber 1995; Baayen 2009, ch.5). In this paper, we undertake such an analysis on the hand-corrected "gold standard" corpus of Russian (1.6 Mio tokens), annotated with morphosyntactic tags and (text-wide) ascriptions of FS and text types (article, critique, discussion, speech etc.) -- `http://ruscorpora.ru/en/corpora-usage.html`. FS-related candidate properties were derived from the relevant literature (28 clearly identifiable features from a set of >600, based, among others, on Kožina et al. 2010). They included (i) lexical occurrences (e.g. particles and adverbs), (ii) certain morphemes (e.g. in internationalisms), (iii) POS categories (e.g. nouns, gerunds). A factor analysis with 3 factors was run on these categories across text types, dimension scores were calculated for factors above a certain threshold, and text types were ranked according to their dimension scores (cf. Biber 1995). Preliminary conclusions are: (i) Factor 1 (main dimension of variation) supports an interpretation like „reporting actions" vs. „static, argumentative". The written/spoken divide does not seem to be at stake here, discussions and conversations being at opposite ends of the scale. This is but one of the relevant distinctions which have been overlooked in the traditional taxonomy of FS. (ii) Traditional FS cannot explain the opposite scalar positions of certain linguistic features – e.g., AdvP (gerunds) and *–acija* internationalisms should both characterize written, especially scientific FS, contrary to their loadings. The bottom-up register distinctions will be compared to the distributions of well-defined linguistic variables, such as subtypes of sentential noun modifiers.

References: • Biber, D. (1995): Dimensions of register variation. • Baayen, H. (2009): Analyzing linguistic data. CUP. • Kožina, M.N. et al. (2011): Stilistika russkogo jazyka. Nauka. • Lapteva, O.A. (2003): Živaja russkaja reč' s teleėkrana. URSS.