Multi-dimensional analysis of Czech. Pilot study

Václav Cvrček, Zuzana Komrsková, David Lukeš, Petra Poukarová, Anna Řehořková and Adrian Jan Zasina (The Czech National Corpus, Charles University, Czech Republic)

Multidimensional analysis (MDA) of register variation (Biber 1991; Biber & Conrad 2009) has proven its worth in the empirical study of English and a typologically varied handful of other languages (incl. Spanish, Chinese, Portuguese, Nukulaelae Tuvaluan, Korean, Somalian). However, it has never been extensively applied to Slavic languages, which are known for their rich inflection, distinctive morphology (e.g. verbal aspect) and a fairly long literary tradition shaping the registers and styles of different genres and text types. This paper aims to discuss specific issues encountered when applying MDA to Czech (a West Slavic language), as well as to point to some methodological innovations we adopted while working on the project.

Although there is sufficient Czech language data (in terms of both extent and diversity) available for this type of research, no extensive investigation of this area has been carried out so far (with the exception of an unfinished and unpublished study by Vilém Kodýtek). Czech is also a highly interesting language for this type of research from the point of view of methodology: it exhibits a high degree of inflection (which goes hand in hand with abundant morphological variation) and a sociolinguistic situation bordering on diglossia (see e.g. Bermel 2014 for a recent overview).

Corpus compilation

The primary goal was to create a corpus as diverse and as representative of the wide range of uses of language as possible. At the topmost level, texts are classified into three modes of communication: written language, spoken language and internet communication. Each mode is further sub-divided into two or more divisions (e.g. the written mode has a fiction, non-fiction, journalism and private correspondence division), divisions then branch further into classes of texts. It needs to be emphasised that the classification mentioned above is not based on intratextual criteria (i.e. on the language used), but on extratextual properties of a text, namely its overall intended function (e.g. poem, scientific paper, column etc.).

The corpus consists of 45 classes of text, each represented by approx. 200,000 words, i.e. 9,074 mil. words in total (excluding punctuation). In order to achieve as diverse a composition of the corpus as possible, we decided to use text samples/chunks (instead of whole texts) of the same length (between 2,000–5,000 words). During the sampling procedure, we took into consideration that the beginnings, middle portions and endings of longer texts should be represented equally. The vast majority of texts in the corpus come from the Czech National Corpus's own resources (see http://wiki.korpus.cz/doku.php/en:cnk:uvod); genres/classes which were not covered by in-house language data production were acquired from other research centres focusing on collecting Czech linguistic data.

For almost all classes, we had more data than we needed (the exceptions being private correspondence and administrative texts). We sampled each class separately (stratified sampling), while paying particular attention to within-stratum diversity.

Features and their operationalization

Once morphologically annotated and lemmatized, the corpus was then searched for more than 140 features, ranging from phonology, morphology and word formation to syntax, lexicon and pragmatics. The list was partially drafted on the basis of an overview of the existing style and grammar literature on individual competing alternatives in Czech (e.g. Čmejrková & Hoffmannová 2011 or Čechová, Krčmová & Minářová 2008, to name but two), and partially on the basis of language variation findings resulting from designing applications for morphological annotation. Unlike other MDA studies, our list of features puts greater emphasis on morphological variation, lexicon-level variation, and type-based features (complementing the commonly used frequency-based characteristics).

The paper will focus on specific problems related to these features:

- frequency of noun cases – morphological features rely heavily on automatic morphosyntactic annotation, which exhibits a significant portion of false positives, especially in spoken and internet texts
- gender markers – the aim is to investigate the biases concerning texts by/about women by examining the variability between texts with different proportions of active female participants (N.B. grammatical gender in nouns is heavily lexicalized in Czech)
- type-based features (inventory of pronouns, prepositions, conjunctions) – as a complement to frequency characteristics of some POS categories, we also decided to include the size of the inventory (number of types) as a separate feature. The higher the number of different items used in a text, the more complex, situationally rooted and/or co-textually linked the text is. The use of uncommon grammatical words within a text of small size (up to 5,000 words) shows a high level of lexical richness.
- lexical classes (abstract nouns, taboo words, poetic words etc.) – problems pertaining to the definition of these classes and their precise delimitation (recall of the feature)
- lexical richness, thematic concentration, use of unigrams and bigrams – esp. the problem of normalization of the scores according to text/chunk length.

Preliminary results

The statistical evaluation of the features as measured on the individual texts is performed via exploratory factor analysis, which enables an interpretation of the multidimensional space (number of texts × number of features) using several two-dimensional scales (for an extended discussion of the methodology in the context of linguistics, see Biber 1991). Each of these can then be made sense of according to the features with which it correlates, e.g. as setting apart texts with a prevailing informative function from texts which are primarily subjective.

As of December 2015, the interim results show a clear separation between spoken and written texts, as expected. According to the features with highest loadings, the first dimension can be interpreted as differentiating dynamic/narrative texts from texts where a descriptive approach prevails; the second dimension captures the difference between unprepared, situationally anchored texts and prepared texts unrelated to the situation of production/reception.

With regards to this top-level differentiation, some inferences can be ventured about the role and position of internet communication which are of utmost importance for the strategies of corpus compilation. When considering the first two dimensions, the web domain seems to overlap with the area of private correspondence (letters), fiction, journalism, non-fiction and (to some extent) also formal spoken communication. However, according to our preliminary results, informal spoken discourse does not seem to be replaceable by internet sources as far as recourse to specific linguistic means is concerned.

References

Bermel, N. (2014). 'Czech Diglossia: Dismantling or Dissolution?' In J. Árokay et al. (Eds.), *Divided Languages?* (pp. 21–37). Wien, Austria: Springer.

Biber, D. (1991). *Variation across speech and writing*. Cambridge, England: Cambridge University Press.

Biber, B., & Conrad, S. (2009). *Register, Genre, and Style*. New York, NY: Cambridge University Press.

Čechová, M., Krčmová, M., & Minářová, E. (Eds). (2008). *Současná stylistika*. Prague, Czech Republic: Nakladatelství Lidové noviny.

Čmejrková, S., & Hoffmannová, J. (Eds). (2011). *Mluvená čeština: Hledání funkčního rozpětí* . Prague, Czech Republic: Academia.