[SLAVICORP]

ORATOR: A new corpus of spoken Czech

A key distinction in language is that between the spoken and written mode. The former is prototypically associated with spontaneous, context-embedded, on-the-fly communication, while the latter is context-reduced, and to minimize misunderstanding, it can and must rely on more careful wording and rewording, often leading to more formal language. While it is useful to keep these prototypes in mind as signposts, in practice, the boundaries can be blurry. Crossing over into "spoken" territory, the niche which immediately comes to mind is real-time communication over the web, where the requirement of responsiveness trumps editing. In the other direction, we have for instance situations where a speaker addresses an audience in an extended monologue.

The 1M-word ORATOR spoken corpus focuses precisely on this second transition area, where spoken language enters "written" territory. It collects monologues by native speakers of Czech, delivered at various occasions which the speakers knew about in advance and therefore could prepare their appearance to a degree they saw fit. The corpus draws a line at speeches which are obviously fully read: the goal is not to examine speech *reproduction*, but *production* under conditions where the speaker has to fill a pre-allocated time slot on a previously agreed upon topic, representing a scientific field, a professional group, or performing a social role – 12 types of situations altogether, from toasts to lectures.

ORATOR complements the landscape of existing corpora of spoken Czech, covering e.g. intimate discourse (ORTOFON, Komrsková et al., 2017) or media speech (DIALOG). This brings Czech closer to languages like English or German, which have publicly available spoken corpora covering a range of different communication situations (e.g. BNC1994 – Aston and Burnard, 1998; or FOLK – Schmidt, 2016).

Aston, G. – Burnard, L. (1998). The BNC Handbook: Exploring the British National Corpus with SARA. Edinburgh: Edinburgh University Press.

Komrsková, Z. - Kopřivová, M. - Lukeš, D. - Poukarová, P. - Goláňová, H. (2017). New Spoken Corpora of Czech: ORTOFON and DIALEKT. Jazykovedný časopis, 68(2), 219-228.

Kaderka, P. – Havlík, M. – Svobodová, Z. – Peterek, N. – Havlová, E. – Klímová, J. – Kubáčková, P. (2008). Minulost, současnost a budoucnost korpusu DIALOG [The past, present, and future of the DIALOG corpus]. In: F. Štícha & M. Fried (eds.), Grammar & Corpora / Gramatika a korpus 2007. Prague: Academia, pp. 181–189.

Schmidt, T. (2016). Good practices in the compilation of FOLK, the research and teaching corpus of spoken German. International Journal of Corpus Linguistics, 21(3), 396–418.