Multi-tier transcription of informal spoken Czech: the ORTOFON corpus approach

Marie Kopřivová, Hana Goláňová, Petra Klimešová, Zuzana Komrsková, David Lukeš

Institute of the Czech National Corpus, Faculty of Arts, Charles University

{marie.koprivova,hana.golanova,...}@ff.cuni.cz

The spoken corpus ORTOFON is currently in the stage of data collection and annotation and will feature two main tiers of transcription: the ort layer (which is more or less orthographical) and the fon layer (which contains a simplified phonetic transcript). The recordings are of the same nature as those in the ORAL series corpora (Kopřivová & Waclawičová 2006; Waclawičová, Křen & Válková 2009): they target prototypical spoken language as instantiated in informal conversations among people who know each other and are situated in their usual environment (at home with their family, among friends, in a restaurant etc.). Our recording associate usually takes part in the dialogue and performs his/her usual role in the group of speakers.

Like previous spoken corpora, ORTOFON will be balanced with respect to several sociolinguistic categories of the included speakers: gender, age, education and dialect region of childhood residence. It will thus allow for interesting comparisons with older dialect recordings (Balhar et al. 2011) which are currently being made into a corpus (called DIALEKT), one of whose layers of transcription will be compatible with the ORTOFON ort layer. Recordings are being collected from all over the Czech Republic, with great emphasis on quality, which is necessary because of the phonetic transcription stage. Apart from face-to-face interactions, telephone or VoIP conversations are also allowed for inclusion.

By offering a detailed multi-tier transcript (including orthographic, phonetic and meta-linguistic layers), we aim to capture interactions in a complex way in the context of a given communication situation. The ort layer is optimized for allowing a reasonably quick first transcription of the sound recording. Being based on orthography, it is mostly intuitive for our non-linguist collaborators and easily searchable. At the same time, it already encodes several phenomena typical of spoken language, e.g. [v]-prothesis, Common Czech endings and dialectal features. The carefully negotiated trade-off between standard spelling and variation makes it possible to track these features' areal distribution in a fairly straightforward way.

More pronunciation details are available via the linked fon layer, which is an innovation compared to the ORAL series. It does not aim to capture all phonetic variation (e.g. vowel quality changes are mostly limited to reduction), but still offers rudimentary pointers to a variety of connected speech processes (Farnecani & Recasens 2010, 322): assimilation of voicing, place or manner of articulation; stress group boundaries; epentheses and elisions etc. Comparison with the ort layer reveals deletions in common words and filler expressions.

Examples will illustrate the specificities of our transcription guidelines, which are currently mostly stable, though still a work in progress in some respects, based on feedback and practical experience. Changes from previous versions will be highlighted as they often offer an interesting perspective on annotation choices.

Keywords: spoken corpus, corpus annotation, transcription, Czech, informal interactions
Linguistic field: Sociolinguistics, Corpus/Computer linguistics

References

Balhar, Jan et al. 2011. *Český jazykový atlas – Dodatky*. Prague: Academia.

Čmejrková, Světla, and Jana Hoffmannová (eds). 2011. *Mluvená čeština: hledání funkčního rozpětí*. Prague: Academia.

Ehlich, Konrad, and Jochen Rehbein. 1976. "Halbinterpretative Arbeitstranskriptionen (HIAT)." *Linguistische Berichte* 45: 21–41.

Farnetani, Edda, and Daniel Recasens. 2010. "Coarticulation and Connected Speech Processes." In *The Handbook of Phonetic Sciences*, edited by William J. Hardcastle, John Laver & Fiona E. Gibbon, 316–352. Chichester, U.K.: Wiley-Blackwell.

Hála, Bohuslav. 1967. *Výslovnost spisovné češtiny I*. Prague: Academia.

Kaderka, Petr, and Zdeňka Svobodová. 2006. "Jak přepisovat audiovizuální záznam rozhovoru? Manuál pro přepisovatele televizních diskusních pořadů." *Jazykovědné aktuality* 43 (3–4): 18–51.

Kopřivová, Marie, and Martina Waclawičová. 2006. "Representativeness of Spoken Corpora on the Example of the New Spoken Corpora of the Czech Language." In *Труды международной конференции "Корпусная лингвистика – 2006"*, 174–181. Санкт-Петербург: Издательство СПбГУ.

Krčmová, Marie. 2010. *Úvod do fonetiky a fonologie pro bohemisty*. Ostrava: Filozofická fakulta OU.

Palková, Zdena. 1994. *Fonetika a fonologie češtiny*. Prague: Karolinum.

Pořízka, Petr. 2009. *Transkripce a sběr dat v korpusech mluvené češtiny*. (Unpublished doctoral dissertation). Filozofická fakulta Univerzity Palackého, Olomouc.

Waclawičová, Martina, Michal Křen, and Lucie Válková. 2009. "Balanced Corpus of Informal Spoken Czech: Compilation, Design and Findings." In *Proceedings of the 10th Annual Conference of the International Speech Communication Association INTERSPEECH 2009*, 1819–1822. Brighton: Curran Associates, Inc.

Zeman, Jiří. 2008. *Základy české ortoepie*. Hradec Králové: Gaudeamus.