A corpus of Czech aphasic speech: development and possible applications

Michal Láznička (michal.laznicka@ff.cuni.cz)

Charles University in Prague

Aphasia presents an important and challenging field of study for linguistics. It can provide important evidence for models of language structure and functioning. On the other hand, linguistic models can influence the work of speech language therapists. One of the problems often encountered in linguistic research of aphasia is the limited availability of relevant data and difficulties regarding data collection. A possible way to surmount this problem is the creation of specialized corpora of aphasic speech, as exemplified by AphasiaBank (MacWhinney et al. 2011), one of the few existing projects, a database of video recordings and transcriptions, based on the CHILDES system (MacWhinney 2000). Aphasic speech and language pathologies in general present one of the most understudied areas within Czech linguistics, only few accounts of the linguistic symptoms exist (Lehečková 2001) and theory driven linguistic analyses are lacking (but see Flanderková et al. 2014). In this paper, I present a project which aims to advance the research in this area by developing a corpus of Czech aphasic speech which will serve as source of data for both linguists and clinicians. The corpus includes 10 hours of structured and semi-spontaneous discourse of 11 individuals with aphasia with different levels of fluency and severity ranging from mild to moderate. The corpus is lemmatized, morphologically tagged, and contains error annotation marking errors typically encountered in aphasic speech, such as paraphasias or agrammatisms. It contains transcripts and time aligned audio tracks. The corpus will be integrated within the Czech National Corpus environment.

To illustrate possible applications of the corpus, I present an analysis of narrative discourse production using a subcorpus of a story retelling task, in which participants saw and retold a three minute video clip. Using several linguistic measures (cf. e.g. Lind et al. 2009), preliminary profiles of Czech aphasic discourse are created, a topic which has not been addressed previously. These measures include type-token ration of verbs and nouns, use of general all-purpose (GAP) verbs and nouns, distribution of case forms, number of pauses, errors, and repairs, frequency profiles of verbs and nouns, number of adjectival and adverbial modifiers, number of clauses per conversational unit, and the proportion of full sentences and sentence fragments. I expect to find a general tendency to use high frequency, GAP items, simpler sentence structure and less background information, individual variation notwithstanding, as suggested in the literature (e.g. MacWhinney et al. 2010).

Word count = 395 words

References:

Flanderková, E., Mertins, B., Bezdíček, O., Baborová, E., Černá, M. (2014) Posuzování gramatičnosti v Brocově afázii - příklad dvou pacientů. *Česká a slovenská neurologie a neurochirurgie 77/110(2)*: 202-09.

Lehečková, H. (2001). Manifestation of aphasic symptoms in Czech. *Journal of Neurolinguistics 14*: 179-208.

Lind, M., Kristoffersen, K. E., Moen, I., Simonsen, H. G. (2009). Semi-spontaneous oral text production: Measurments in clinical practice. *Clinical Linguistics & Phonetics 23(12)*: 872-886.
MacWhinney, B. (2000). *The CHILDES Project: Tools for Analysing Talk (3rd ed.)*. Mahwah, NJ: Lawrence Erlbaum Associates Inc.
MacWhinney, B., Fromm, D., Holland, A., Forbes, M., Wright, H. (2010). Automated analysis of the Cinderella story. *Aphasiology 24*: 856-868.
MacWhinney, B., Fromm, D., Forbes, M., Holland, A. (2011). AphasiaBank: Methods for Studying Discourse. *Aphasiology 25*: 1286-1307.