A Learner Corpus of Spoken Danish and its Direct and Indirect Applications in Language Teaching

Thomas MATHIASEN & Mikołaj SOBKOWIAK
Department of Scandinavian Studies
Adam Mickiewicz University
Poznań, Poland
tommat@amu.edu.pl; miksobko@amu.edu.pl

The main purpose of this paper is to present how a learner corpus of spoken Danish can and is being used in teaching Danish as a foreign language. The authors show that the methods applied in data acquisition and tagging make the corpus in question both a practical tool for linguistic analysis and a useful resource for language teachers and students.

Barely any research has been done (and none in Poland) on the acquisition of Danish as a foreign language by Poles despite the fact that over 150 students are currently studying Danish language and culture at 4 major Polish universities, hundreds of Poles learn Danish at private language schools and tens of thousands of Poles work in Denmark and learn the language there.

The acquisition of corpus data started in October 2011 and is scheduled to continue until June 2014. Ultimately the corpus will comprise of audio recordings of *clinically-elicitated* (cf. Ellis & Barkhuizen, 2005) individual utterances and language interactions of groups of Polish students of Danish language and culture. It will follow their acquisition of Danish over a 3-year course of intensive study (between 8 and 12 hours a week). During this period the students will acquire a command of the Danish language equivalent to the B2/C1 level in the Common European Framework of Reference for Languages. The acquired linguistic data will be fully transcribed (using the CLAN software provided by CHILDES) and tagged, using a system of morphosyntactic annotation (a so-called MOR grammar) for Danish. Both the corpus and the annotation system will be released into an Open Access database (TalkBank).

An important part of the corpus is its error annotation system. The tagging system has been designed and continues to be developed by the authors of this paper, and is compatible with the CLAN speech transcription and analysis software. Error annotation within the system is based on two main error categories, i.e. phonological and grammatical errors, the latter being further divided into subclasses concerning the lexicon, word order, inflection, definiteness, agreement, concord, case and overall comprehensibility and others. The mentioned error subclasses are further divided into over 60 error codes.

When applied to the CLAN program, the developed error codes enable a detailed and comprehensive error analysis of individual recordings as well as across files. Based on the error codes the CLAN program provides complex search possibilities, the scope of the search being a single speaker, any defined group of speakers or any error type(s) defined.

In this way the authors are able to apply the corpus directly and give thorough and detailed feedback to the participants of the study after each recording. The feedback process includes teacher-learner interaction and learner-corpus interaction. Both of these are performed using transcribed recordings synchronized with corresponding audio files, along with error report sheets containing lists of the participants' errors based on error type.

The corpus can also be applied indirectly in teaching. The results of a preliminary study carried out by the authors, using the same methods and tools as for data acquisition for the corpus, have had an impact on the Danish course syllabi at the department where the authors are employed. The authors have also used the findings of the preliminary experiment to further develop a course of Danish phonetics. The acquired corpus data are expected to have a substantially greater impact on the studies of the acquisition of Danish by Poles and, subsequently, on teaching Danish as a foreign language.

References

Beal, J., Corrigan, K. & Moisl, H. (2007). *Creating and Digitizing Language Corpora. Volume 1: Synchronic Databases*. Houndmills, New York: Palgrave Macmillan.

Behrens, H. (ed.). (2008). *Corpora in Language Acquisition Research*. Amsterdam/Philadelphia: John Bejnamins.

Ellis, R. & Barkhuizen, G. (2005). *Analysing Learner Language*. Oxford: Oxford University Press.

Gries, S., Wulff, S. & Davies, M. (2010). *Corpus-linguistic applications. Current studies, new directions.* Amsterdam/NewYork: Rodopi.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*. 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. (2011). *The CHILDES Project: Tools for Analyzing Talk. Part 1: The CHAT Transcription Format.* Available from: http://childes.psy.cmu.edu/manuals/chat.pdf.

MacWhinney, B. (2011). *The CHILDES Project: Tools for Analyzing Talk. Part 2: The CLAN Programs*. Electronic Edition: http://childes.psy.cmu.edu/manuals/clan.pdf.

Sinclair, J. McH. (ed.). (2004). *How to Use Corpora in Language Teaching*. Amsterdam/Philadelphia: John Benjamins.