Towards a Best Practice for sharing and reusing learner corpora: the Hamburg Map Task Corpus

Hanna Hedeland, Timm Lehmberg, Thomas Schmidt, Kai Wörner

The Hamburg Map Task Corpus (HAMATAC) is a small spoken language corpus documenting the performance of advanced learners of German. It is based on a map task experiment designed by the project "Deutsch Heute" at the Institute for German Language (Brinckmann / Kleiner / Knöbl / Berend 2008). A map task is a cooperative task in which one participant has to explain a route marked on his map to a second participant who has a similar (but not identical) map without the route. The design has been used – with different maps and for different languages – in quite a number of corpus studies (e.g. Anderson et al. 1991). It is especially suitable for producing comparable data sets in which certain linguistic constructions (e.g. prepositional phrases) occur particularly often.

After presenting the general design and construction workflow of the Hamburg Map Task Corpus as well as an exemplary analysis, we will use it as a starting point for discussing different aspects of a best practice for making learner corpora reusable and sharable, such as:

- Comparibility and reproducability of corpus design: as mentioned above, HAMATAC's design has been used in identical form for a corpus of German native speakers, and similar designs have been used for corpora of other languages. These points of contact make it easier for researchers to compare their findings on the corpus to results gained with other data. Furthermore, the design is easily reproducible so that additional comparable data (e.g. with less advanced learners of German or advanced learners of another language) can be easily obtained.
- Data protection issues: HAMATAC participants were asked to sign an agreement that the resulting data can be made available for research purposes. This is an indispensable, but often overlooked, prerequisite for sharing the corpus.
- Theory-dependence and quality of annotations: to make a corpus reusable for a wider community with heterogeneous theoretical backgrounds and interests, it is crucial that those annotations which are uncontroversial with respect to different theoretical approaches are clearly separated from annotations which are owing to a specific theoretical assumption. Often, it is only the first type of annotation that is readily reused by other researchers. Issues of quality control for this type are therefore particularly important.
- Interoperability and standardisation of formats: HAMATAC was created with the help of the EXMARaLDA system (Schmidt/Wörner 2009). The resulting data, however, are made available in a variety of formats so that researchers can reuse and analyse it with the tool(s) of their choice. EXMARaLDA's interoperability functionality makes sure that the most important tool formats (e.g. ELAN or Praat) and standards (e.g. TEI or AG) are covered.

Anderson, A./Bader, M./Bard, E./Boyle, E./Doherty, G. M./Garrod, S./Isard, S./Kowtko, J./McAllister, J./Miller, J./Sotillo, C./Thompson, H. S./Weinert, R. (1991): The HCRC Map Task Corpus. Language and Speech, 34, pp. 351-366.

Brinckmann, Caren/Kleiner, Stefan/Knöbl, Ralf/Berend, Nina (2008): German Today: an areally extensive corpus of spoken Standard German. In: Proceedings 6th International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Marokko.

Schmidt, T. & Wörner, K. (2009): EXMARaLDA – Creating, analysing and sharing spoken language corpora for pragmatic research. In: Pragmatics 19.