## From project data to sustainable archiving of linguistic corpora

*Christian Chiarcos, Erhard Hinrichs, Georg Rehm, Thomas Schmidt and Andreas Witt*

This paper describes a new research initiative addressing the issue of sustainability of linguistic resources. This initiative is a cooperation between three collaborative research centers in Germany which comprise more than 40 individual research projects altogether.

Each of these research centers features one central "data" project whose main task it is to develop methods for enabling or facilitating the exchange and reuse of empirical linguistic data between individual projects. As is the case with much linguistic data, such an exchange or reuse tends to be difficult because individual data sets strongly depend on specific technological environments and on specific theoretical backgrounds. To overcome these difficulties, XML-based frameworks were developed at all three sites that allow researchers to represent diverse types of written or transcribed spoken language corpora on a common structural basis. These frameworks are: TUSNELDA (developed at the SFB 441 in Tübingen), EXMARaLDA (SFB 538, Hamburg) and PAULA (SFB 632, Potsdam). With these frameworks, some of the most fundamental obstacles in handling linguistic corpora can be overcome. The aim of the new project described here is to address further issues that must be solved on the way towards truly sustainable archives of linguistic data. This will involve work in seven areas:

1) Annotation frameworks: We plan to develop an annotation framework generalizing over the three existing ones and thus usable as a sustainable archiving format for all corpora. Most importantly, this will mean finding ways of bringing together time-centric and hierarchy-centric conceptions of linguistic data. The NITE Object Model is chosen as a starting point for this.

2) and 3) Query tools and other tools for data access: Appropriate ways of finding and querying archived data are among the crucial prerequisites for sustainable data handling. Work in this area will focus on developing adequate query languages and user interfaces for linguists with diverse backgrounds.

4) Meta data: As corpora are made available to a wider range of users, their documentation in the form of meta data becomes more important. Work in this area will depart from existing suggestions like IMDI and OLAC.

5) Ontologies: Linguistic ontologies can acts as a further aid in exchanging and reusing linguistic data. The GOLD ontology is chosen as a starting point in this area.

6) Rules of best practice: In order to ensure that future corpus creators can profit from existing experience, we plan to publish a set of best practice rules for the creation and dissemination of linguistic corpora.

7) Legal and ethical questions of data handling: Legal or ethical issues can play a very important role in the dissemination of linguistic corpora. This work package will address these issues, again by aiming at a set of best practice rules to be made available to the research community.

By presenting problems solved in previous work at the different sites and problems yet to be tackled in this new joint project, this paper will give an overview of what we see as the state of the art in digital language documentation.