Data structures for the analysis of regional language variation

Birgit Kellner, Timm Lehmberg, Ingrid Schröder and Kai Wörner

Abstract. This article reports on work in progress on the development of data structures and processing methods which take into account the special demands for the documentation and analysis of regional language variation. This work is an integral part of the supra–regional project "Language Variation in Northern Germany" which started in the beginning of 2008 as a joint initiative of six Northern German universities.

1 Introduction

The project "Language Variation in Northern Germany" ("Sprachvariation in Norddeutschland" — SiN) aims at the documentation, description and analysis of the usage of dialectal variation in Northern Germany along a spectrum between Low German and High German.

This objective requires the analysis of object language material from communication settings with different degrees of formality as well as metalinguistic information on speaker biographies, language awareness and language attitudes. The data used for this purpose is gathered through various settings of acquisition (translation, interview, spontaneous talk, salience tests etc.) and thus is highly heterogeneous and complex with regard to the respective level of dialectological analysis. It is the basis of a deeply annotated over–all Northern German corpus of spoken language. The implementation of such a corpus forces the use of sophisticated data standards and tools for the processing (recording, transcriptions, annotation etc.) and analysis of spoken language.

The article is structured as follows: Section 1 gives an overview of the research problem as well as the chronological and geographical structure of the project. Based on this, the methods and principles of data acquisition are presented. Section 2 shows the present state of the implementation with assistance of the EXMARaLDA system which is used for the processing of the entire language data. A special focus is laid upon the modelling of the metadata and the problems of a multi–layer annotation of transcribed spoken language.