Preface

The papers in this volume were presented at the workshop *Heterogeneity in Linguistic Databases*, which took place on July 9, 2004 at Potsdam University. The workshop was organized by project D1: *Linguistic Database for Information Structure: Annotation and Retrieval*, a member project of the SFB 632, a collaborative research center entitled *Information Structure: the Linguistic Means for Structuring Utterances, Sentences and Texts*.\(^1\)

A prominent feature of the SFB 632 is that it unites projects that base their research on empirical data. The individual projects collect and annotate linguistic data of various types, which then constitutes the ground for further research. A central project provides the technical infrastructure for building, maintaining, and retrieving this data. Similar to many current comparable research projects, the SFB deals with very heterogeneous linguistic data. This heterogeneity results from a number of factors: First, primary data itself is heterogeneous, differing with respect to size (e.g., single sentences vs. sentences in context), modality (monologue vs. dialogue, text vs. speech), and language. Second, rich, multi-level annotations often require data structures of various types (attribute-value pairs, trees, pointers, etc.). Third, data is often annotated by means of different, task-specific annotation tools (e.g., by tools for syntax, discourse, or co-reference annotation). Furthermore, similar to the data, the needs and backgrounds of potential users of the data vary: Researchers come with tasks of diverse types—ranging from manual exploration to automatic statistical computations—and differ with regard to their computer skills. This diversity must be reflected by the retrieval facilities.

The workshop brought together developers and users of linguistic databases from a number of research projects that work on an empirical basis, all of which have to cope with some of the issues sketched out above. The first four papers address aspects of heterogeneous data from the point of view of database developers; the remaining three papers focus on data exploitation by the users.

In his paper *Unity in Diversity: Integrating Differing Linguistic Data in TUSNELDA*, Andreas Wagner presents the TUSNELDA corpus, a collection of diverse sub-corpora, which differ with regard to object languages, text types, text types,\(^1\)

\(^1\)For more information about the SFB 632, visit [http://www.sfb632.uni-potsdam.de/](http://www.sfb632.uni-potsdam.de/), for project D1, see [http://www.sfb632.uni-potsdam.de/projects/d1](http://www.sfb632.uni-potsdam.de/projects/d1).
annotation types, and underlying linguistic theories. The underlying, XML-based annotation scheme TUSNELDA is both sufficiently flexible to cover the diversity of the data as well as uniform enough to enhance data (re)usability.

Thomas Schmidt’s paper *EXMARaLDA und Datenbank, Mehrsprachigkeit* — *Konzepte und praktische Erfahrungen* (in German, with extended abstract in English) addresses concepts and principles in the development of a database for heterogeneous speech data, such as the importance of standardization at different levels of data processing. He also reports practical experiences with users and technologies, including time costs in developing the database.

In her paper *Heterogeneity and Standardization in Data, Use, and Annotation: a Diachronic Corpus of German*, Anke Lüdeling describes problems specific to diachronic corpora, such as the highly variable orthography of old texts. She proposes a flexible corpus design that encompasses this variability and, at the same time, supports standardized annotation.

*Multiple Hierarchies: New Aspects of an Old Solution* by Andreas Witt addresses problems posed by data annotations with (i) multiple/overlapping hierarchies and (ii) heterogeneous tagsets. He discusses different SGML/XML-based solutions and proposes a redundant encoding, which replicates the primary data for each hierarchy and tagset and makes use of Prolog facts for data representation.

Roland Meyer’s paper *VP-Fronting in Czech and Polish—A Case Study in Corpus-Oriented Grammar Research* shows how corpus data can be used for the examination of a specific linguistic phenomenon. He examines VP-fronting in Czech and Polish and its information-structural facets. Referring to different corpora with varying types of annotation, he discusses user requirements with regard to annotations and search facilities.

George Smith (*Refining Queries on a Treebank with XSLT Filters. Approaching the Universal Quantifier*) argues for the importance of the universal quantifier in query languages, which is a prerequisite for (easy) retrieval of many linguistic phenomena but is often not implemented due to efficiency reasons. He presents XSLT stylesheets that implement the universal quantifier; these stylesheets can be used, for example, to further filter the results of a TIGERSearch query.
In their paper *Exploring Lexical Patterns in Text: Lexical Cohesion Analysis with WordNet*, Elke Teich and Peter Fankhauser present an analysis and exploration of lexical cohesion. First, lexical chains are annotated in a corpus, based on WordNet. A specific tool then allows the user to analyze the chains manually. Statistic analyses of the data show, among other things, that length and type of lexical chains depend on the register of the text.

We are very grateful to the authors for their contributions to this volume. We would also like to thank them for their presentations at the workshop, which stimulated some very interesting and fruitful discussions.

Stefanie Dipper
Michael Götze
Manfred Stede
Contents

Unity in Diversity: Integrating Differing Linguistic Data in TUSNELDA  
Andreas Wagner ................................................................. 1

EXMARaLDA und Datenbank ‘Mehrsprachigkeit’ — Konzepte und praktische Erfahrungen  
Thomas Schmidt .............................................................. 21

Heterogeneity and Standardization in Data, Use, and Annotation: a Diachronic Corpus of German  
Anke Lüdeling ................................................................. 43

Multiple Hierarchies: New Aspects of an Old Solution  
Andreas Witt ................................................................. 55

VP-Fronting in Czech and Polish—A Case Study in Corpus-Oriented Grammar Research  
Roland Meyer ................................................................. 87

Refining Queries on a Treebank with XSLT Filters. Approaching the Universal Quantifier  
George Smith ................................................................. 117

Exploring Lexical Patterns in Text: Lexical Cohesion Analysis with WordNet  
Elke Teich and Peter Fankhauser ...................................... 131