INTRODUCTION

JANNE BONDI JOHANNESSEN

Department of Linguistics and Nordic Studies, University of Oslo

Research infrastructure has become more and more important in recent years, not just in the natural sciences, but also in the humanities and especially in linguistics. While linguists over the last fifty years have often found it sufficient to use their own intuitions about grammaticality by way of introspection when making claims about language structure, such an approach is rarely seen as acceptable in the 21st century. With electronic text corpora and even speech corpora the linguists have far better opportunities for making sound claims about empirical matters, and, importantly, in a way that can be scientifically checked by other researchers. In the study of language variation, it is even more important to have access to resources that document this variation.

This volume was initiated after a workshop on Research Infrastructure for Language Variation Studies (RILIVS), held at the University of Oslo in the autumn of 2009. The workshop was the main one of a series of three exploratory workshops funded by the Nordic research fund NOS-HS and initiated by researcher Øystein Alexander Vangsnes (University of Tromsø) with professor Janne Bondi Johannessen (University of Oslo), in addition to the following people: professor Odd Einar Haugen (University of Bergen), professor Anders Eriksson (University of Gothenburg), professor Höskuldur Thráinsson (University of Iceland), professor Jan-Ola Östman (University of Helsinki), assistant professor Karen Margrethe Pedersen (University of Copenhagen), assistant professor Henrik Jørgensen (University of Aarhus), professor Christer Platzack (University of Lund), researcher Bruce Morén (University of Tromsø), professor Tor A. Åfarli (Norwegian University of Science and Technology). The major aims of the workshop can be summarised thus:

- Bring together scholars and technicians from various projects concerning linguistic variation in order to:
 - consolidate the resources developed in the individual projects
 - establish common and realistic research goals in the field of linguistic variation
 - develop common standards and tools to enhance the quality of linguistic variation research

- Goals to be identified should be targetted at the needs and desires of the linguists and not delimited by the current state of technology. Taking this general stand will foster innovativeness rather than mere application of existing resources.

This volume consists of papers on topics of essentially two types: on the use of language research infrastructure and on development of such infrastructure.

Patrick Bye is interested in linguistic innovations and describes how mapping these geographically can help reconstructing the history of certain linguistic features. He argues that for a particular linguistic feature it may be difficult to determine the order of development just by using simplified generalisations. Instead, important insight can be gained by looking at the actual geographical distribution of a feature, taking into account factors such as difficult terrain, knowledge about the history of other linguistic features, and also the cultural, archeological and historical context of the relevant geographical area.

Harald Hammarström and *Sebastian Nordhoff* focus on the problem of proliferation of linguistic documents and references. In particular, they describe *LangDoc*, a project for the creation of a website for bibliographical information on linguistic typology. There can be no doubt as to the usefulness of this kind of annotated database, since existing ones are either skewed w.r.t. the authors or types of literature that they represent, or cover only a subset of the languages of the world, or are expensive.

Janne Bondi Johannessen's paper presents the Nordic Dialect Corpus, which is a joint infrastructure where linguists from all the Nordic countries have contributed with dialect recordings, which have subsequently been transcribed, and annotated and made available in a user-friendly searchable corpus. The corpus thus consists of spoken texts from five languages, and is searchable w.r.t. words and parts of words, annotations, phonetic form, and informant metadata. The data are presented in a multimedia display for transcriptions, audio and video, and there are many further options for frequency, sorting, translations and even distribution of search results on geographical maps.

Jan Pieter Kunst and *Franca Wesseling* describe the European Edisyn project, which has a two-fold linguistic goal: on the one hand to establish a network of dialect linguists who have the same standards w.r.t. methodology and data collection, and on the other to study a particular phenomenon: doubling. However, in order to be able to better study the linguistic variation in the languages affiliated with the project, an important outcome is the development of a common database: The Edisyn Search Engine. The database so far consists of five corpora, covering Dutch, Estonian, Northern Italian, Portuguese, and Scandinavian dialects (incidentally the Nordic Dialect Corpus, which is also represented in this anthology). The prospect of a common search engine across languages as differ-

ent as those mentioned is an exciting one for future linguistic research.

Therese Leinonen focusses on tools for illustrating dialect variation, and in particular, dialect levelling in Swedish. She has used phonetic dialect data available from a previous dialect project, and put them through first acoustic analysis and then dialectometrical methods, in particular aggregating techniques like cluster analysis and multidimensional scaling. The various maps that are projected on the basis of the calculations are very good for illustrating the status of the variation. By using variables such as the age of informants in addition to geographical site it is possible to see how a dialect is changing even if the data have all been recorded at the same time. The changes are very obvious once they are projected in the whole colour spectrum.

Eiríkur Rögnvaldsson, Anton Karl Ingasson and Einar Freyr Sigurðsson present an ongoing project which aims at finding methods for developing a Icelandic treebank. Their paper stresses the importance of using open source software, especially for languages like Icelandic, which can be considered to be less-resourced. Their project benefits from the open source tool-kit IceNLP, which has two kinds of morphological taggers, a shallow parser, a lemmatiser, and a tokeniser. Combined with the two available programs Corpus Search and CorpusDraw, the way to a treebank is relatively short.

Diana Santos presents the Linguateca infrastructure for Portuguese. Linguateca is a Portuguese project that has existed for many years. The resources have not been presented before in full for an international audience, and this is the main purpose of this paper. First of all there is a cluster of corpora containing various text genres, written in Portuguese from three continents. The texts have been grammatically tagged and parsed. Parts of the resources have also been semantically classified using automatic methods.

Xavier Villalba's paper is on the Hispacat database of syntactic constructions and how to apply it to syntactic research. This database contains data suitable for micro-comparative variation studies, and contains both Catalan and Spanish data. Importantly, in contrast to an ordinary text corpus, it also contains negative data. The database has three main goals: to be an empirical database for bilingualism and L2 learning, to provide a comparative grammar, and to contain a catalogue of examples.

Without proper peer reviewing academic work would suffer. We are therefore very grateful to the reviewers who accepted to do this job for us:

Sturla Berg-Olsen, Antonio Fabregas, Jan Terje Faarlund, Kristin Hagen, Janne Bondi Johannessen, Karl-Gunnar Johansson, Terje Lohndal, and Andrew Nevins.

WORKSHOP ON RESEARCH INFRASTRUCTURE FOR LINGUISTIC VARIATION:

<http://www.tekstlab.uio.no/rilivs/>

Speakers and participants at the RILIVS workshop at the University of Oslo 2009.
(Photo: Ram E. Gupta)