The Nordic Dialect Database: Mapping Microsyntactic Variation in the Scandinavian Languages

Arne Martinus Lindstad
University of Oslo
Oslo, Norway
a.m.lindstad@iln.uio.no

Anders Nøklestad
University of Oslo
Oslo, Norway
a.noklestad@iln.uio.no

Janne Bondi Johannessen
University of Oslo
Oslo, Norway
j.b.johannessen@iln.uio.no

Øystein A. Vangsnes
University of Tromsø
Tromsø, Norway
oystein.vangsnes@hum.uit.no

Abstract

We describe the development of a database containing informant judgments on a range of test sentences. The database is intended as a research resource for linguists interested in morphosyntactic variation across Scandinavian dialects. We present the data types contained in the base, and how they are used to create a user-friendly search interface. The database forms part of the efforts undertaken under the ScanDiaSyn project umbrella, currently run at ten universities in Denmark, The Faroe Islands, Finland, Iceland, Norway and Sweden. The database has been developed by the Text Laboratory at the University of Oslo, Norway.

1 Introduction

The Nordic Dialect Database is part of the achievements of the Scandinavian Dialect Syntax (ScanDiaSyn) project umbrella. ScanDiaSyn is a collaborative effort run by individual research groups at ten universities in the Nordic countries. The main purpose of ScanDiaSyn is to chart and study morphological and syntactic variation in Scandinavian dialects. The outcome of the project will be a pan-Scandinavian dialect research resource, made available to the research community via a user-friendly web interface. The data collected for the project are of three kinds:

- Speaker intuitions, i.e. speakers' evaluation of test sentences presented to them in a questionnaire.
- A corpus of transcribed audio and video recordings of interviews of and conversations between the informants.
- "Translation" of constructions into dialect from the standard language.

In this paper, we focus on the speaker intuition data. First, we sketch some background in section 2, then discuss the data types that form the basis for the database in section 3, before showing how the data is made available and searchable via a web resource in section 4. Section 5 briefly presents technical aspects of the database, and section 6 discusses future improvements to the system not yet implemented.

2 Background

Somewhat unevenly distributed across the countries, ScanDiaSyn has gathered data at 270 measure points in Scandinavia.

The data from the questionnaire part of the project forms the basis for the database we have built. A subset from a common pool of around 1400 sentences is tested at each measure point. In Norway, 140 sentences are tested, while in Denmark up to 240 sentences are tested at each point. It is up to each research group to decide exactly which sentences are tested, based on individual interest and on what is considered relevant in each dialect.

Kristiina Jokinen and Eckhard Bick (Eds.)
NODALIDA 2009 Conference Proceedings, pp. 283–286

Though the number of sentences tested is not very high, it is demanding for the informants, as evaluating grammaticality is an unusual task for most speakers.

The database developed so far is based primarily on data from the Norwegian and Danish parts of the project. Data from the other languages will be added when they are available.

3 Data types

Compared to the spoken language data in the corpus (see section 1), the amount of data comprising the database is relatively small, and not very much preprocessing is required. In this section, we describe the various data types that enter into the database.

3.1 Test sentences and constructions

The data collection for the database is inspired by a generative syntax approach to grammatical variation (in terms of parameters). Test sentences are constructed to reflect well-known patterns of variation described in the literature, or they are based on expected patterns of syntactic variation across the dialects.[1]

3.2 Speaker evaluations

Following standard practice within generative linguistics (Chomsky 1965), speaker intuitions (or judgments) on the grammaticality of syntactic constructions are considered crucial for a comprehensive theory of language. Informants are asked to judge test sentences on a five-point scale, where 1 is bad and 5 is fully acceptable.

3.3 Linguistic categorisation

Each test sentence has been appended with a number of linguistic features – or categories – describing in as much detail as possible the linguistic property that is tested by that particular sentence. An illustration is given in (1) and (2), *wh*-questions differing in the placement of the finite verb:

(1) Hva du heter?
what you is.called
'What is your name?'

(2) Hva heter du?
what is.called you
'What is your name?'

[1] Note that the informants never see the test sentences visually. We "translate" each test sentence into the local dialect and record a local speaker reading them aloud. The sentences are then presented to the informants aurally.

The linguistic categories appended to these example sentences are the following:

(3) word order, interrogative, question, constituent question, simple *wh*-word

In addition, a category describing the placement of the finite verb distinguishes the sentences from each other: "V3" for (1) and "V2" for (2).

3.4 Metadata: Demographic information

In the Norwegian subproject, the number of informants per measure point is four, one of each sex below the age of 30, and one of each sex above the age of 50. Following traditional sociolinguistic practice, various types of demographic information about the informants are gathered before the recordings are undertaken. This is described in more detail in section 4.

The charting of demographic information and linguistic background ensures that the individual informant is a genuine speaker of the dialect in question.

4 The user interface

As mentioned in section 3, the amount of data is rather small. The challenge lies in structuring, displaying and making the actual content available to researchers in a user-friendly fashion. Various criteria and variables can be applied for performing searches in the database. Figure 1 is a screen dump of the search interface, illustrating the search possibilities. In this section we describe the search possibilities in detail.

4.1 Main search options: categories and test sentences

For most syntacticians, a search for a given feature in a dialect will typically be based on a special interest in a particular syntactic phenomenon such as variation in the placement of the finite verb in constituent questions (*wh*-questions), as above. This is a phenomenon that splits the Norwegian dialect continuum into regions (cf. Vangsnes (2005) for an overview and further literature on the subject). In Figure 1, a search with categories has been performed. This is done by activating category search in the upper left box of the screen. Categories are listed in the drop-down menu at the top of this box. Selecting a given category pops up a sub-menu with all other categories appearing together with the selected category in the description of any sentence

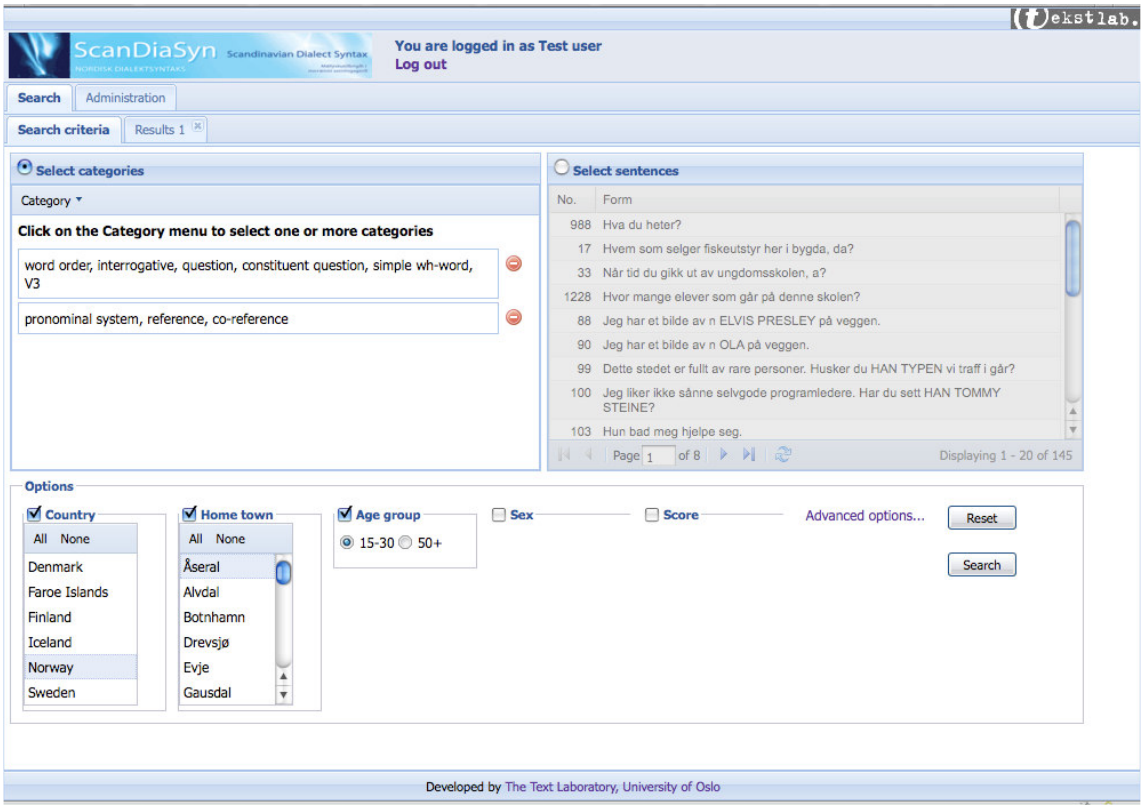

Figure 1: Search interface.

in the database. This way, the search is narrowed, and returns a smaller set of sentences. Several category searches can be specified simultaneously, enabling listing of covariance between phenomena.

This is also illustrated in Figure 1: the user specifies two sets of categories (search criteria), each of which is defined by a comma-separated list. Each set of categories returns a set of one or more test sentences, and the final search result is the union of these sentence sets.

As a second option, the database is searchable by test sentence, i.e., a single sentence or a set of sentences can be selected in the upper right box.

4.2 Restricting the search

While it is possible to search for all judgments for a given test sentence regardless of any variables, it will sometimes be useful to narrow down the search in various ways to obtain a manageable output. This is obviously so if one is looking for covariance between phenomena.

The search can be restricted using the information provided by the various data types described in section 3. In the search interface (Figure 1), this can be accomplished by using the five drop-down menus at the lower end of the screen.

Leftmost, the search can be restricted geographically to a single country or to a combination of countries. This narrows down the set of measure points in the next menu. *Norway* is selected above, and a list of all measure points in Norway is provided in the next menu. Any combination of measure points can be selected for comparison on the features specified in the category search, or on the particular sentences selected in a test sentence search.

If there is agreement between the informants on a particular phenomenon, one can say something meaningful about the dialect in question. Irrespective of dialectal variation, one can also compare the language of e.g. men and women or of young and old speakers over a user-defined geographical area. This is accomplished by specifying the age group and/or the sex in the relevant drop-down menus. For illustrative purposes, the age group 15-30 is selected in Figure 1.

Finally, in the rightmost drop-down menu it is possible to restrict the selection to those sentences that have been given specific scores by the informants, e.g. high acceptance scores, such as 4 and 5 (see section 3.2).

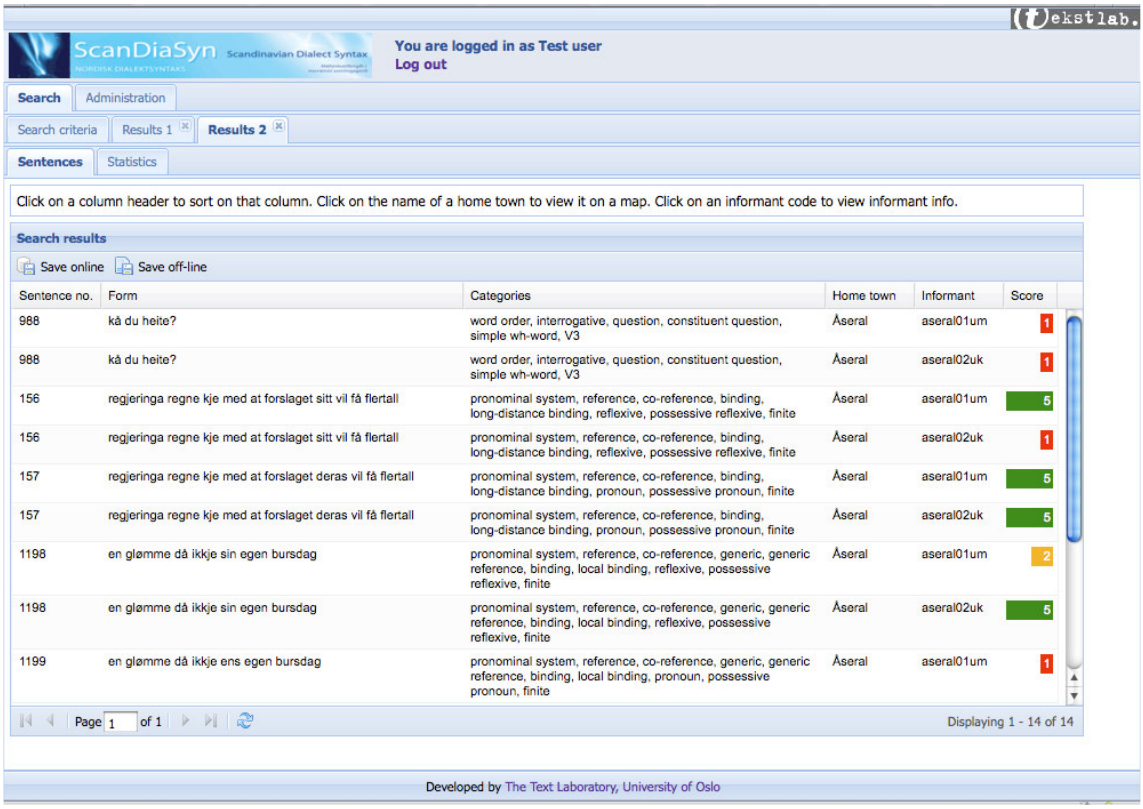

Figure 2: Results page.

4.3 Displaying the results

The results from a given search are displayed in a new tab next to the "Search criteria" tab. Each new search opens a new tab (cf. "Results 1" and "Results 2" in Figure 2). A search can be saved on- or off-line for further processing. Search results are abandoned by closing the tab.

The search results can be sorted in various ways by clicking column headers in the results page, a measure point can be displayed on a map by clicking its name, and demographic information about the informants can be obtained by clicking the informant code.

Throughout, our efforts have been aimed at creating a user-friendly system that can easily adjust to the needs of linguists of any theoretical orientation, and the system is open for easy addition of further variables and search criteria.

5 Technical issues

The server side of the system runs on the Ruby on Rails web application framework[2] with a MySQL database.[3] The web browser interface has been created using the Ext JS JavaScript framework.[4]

6 Refinements: maps and statistics

As a refinement in the future, and for the ease of the eye, a map function will be implemented that can illustrate the presence of a linguistic feature at given places in the dialect continuum. This will enable drawing of isoglosses. Given the dynamic search possibilities the system provides, any covariance between linguistic properties (features, categories) can be easily illustrated in a graphic fashion. We are also planning to provide statistical measures that can be used to detect significant patterns of dialect variation.

References

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, Massachusetts.

ScanDiaSyn: http://uit.no/scandiasyn

The Text Laboratory: http://www.hf.uio.no/tekstlab

Vangsnes, Øystein Alexander. 2005. Microparameters for Norwegian *wh*-grammars. *Linguistic Variation Yearbook*, 5: 187-226.

[2] http://rubyonrails.org
[3] http://www.mysql.com
[4] http://extjs.com

ISSN 1736-6305 Vol. 4
http://hdl.handle.net/10062/9206