Enhancing language resources with maps

Janne Bondi Johannessen, Kristin Hagen, Anders Nøklestad, and Joel Priestley
The Text Laboratory, ILN, University of Oslo
P.O.Box 1102 Blindern. 0317 Oslo, Norway
E-mail: jannebj@iln.uio.no, kristiha@iln.uio.no, noklesta@iln.uio.no, joeljp@iln.uio.no

Abstract

We will look at how maps can be integrated in research resources, such as language databases and language corpora. By using maps, search results can be illustrated in a way that immediately gives the user information that words or numbers on their own would not give. We will illustrate with two different resources, into which we have now added a Google Maps application: The Nordic Dialect Corpus (Johannessen et al. 2009) and The Nordic Syntactic Judgments Database (Lindstad et al. 2009). We have integrated Google Maps into these applications. The database contains some hundred syntactic test sentences that have been evaluated by four speakers in more than hundred locations in Norway and Sweden. Searching for the evaluations of a particular sentence gives a list of several hundred judgments, which are difficult for a human researcher to assess. With the map option, isoglosses are immediately visible. We show in the paper that both with the maps depicting corpus hits and with the maps depicting database results, the map visualizations actually show clear geographical differences that would be very difficult to spot just by reading concordance lines or database tables.

1. Introduction

Creating corpora and databases for linguistic research is an ongoing effort with two almost conflicting goals. On the one hand, the users (linguists and philologists) want as much data as possible, i.e. the more words in the corpus, the better, and the more meta-variables the better. The users want as many different search options and combinations as possible. On the other hand, such resources are hard to combine with another wish by the same users: maximum user-friendliness.

In this paper we will look at how maps can be used to fulfill both options; they add information for the users, and make the resources easier to use. We will illustrate with two different resources, into which we are now adding maps: The Nordic Dialect Corpus (Johannessen et al. 2009a) and The Nordic Syntactic Judgments Database (Lindstad et al. 2009). We have integrated Google Maps into these applications.

2. Nordic Dialect Corpus

The Nordic Dialect Corpus (Johannessen et al. 2009a) is the result of collaboration in the research networks Scandinavian Dialect Syntax and Nordic Centre of Excellence in Microcomparative Syntax. The technical development is carried out at the Text Laboratory at the University of Oslo.

The corpus is under development, in the sense that the number of words is still growing and new functionalities are still being added, but it is fully usable. At the moment it contains 1.5 million words.

The corpus contains recordings of dialects from the more or less mutually intelligible languages of the five countries Denmark, Iceland, Faroe Islands, Norway and Sweden. Recordings have mostly been done on a national basis, which means that there is some variation between them. For example, the Swedish material was mainly recorded during a different project a decade ago, while the Danish and Norwegian material were mainly recorded by national projects with this particular corpus in mind. The Faroese recordings were done as part of the dialect project, while the Icelandic recordings have been done partly in the project and partly before. In addition, while all the languages have been represented by modern language (from the last decade), there are also some older recordings from Norway. All the recordings in the corpus contain spontaneous speech, but while for some of the languages there are conversations between informants, for others, the conversations are between one informant and one project assistant, and for some there are both types. The number of dialects recorded in each country varies due to differences in financing and of course in the linguistic situation. For example, there are around ten in Denmark and will be around 100 in Norway and Sweden.

The recordings are presented in audio and, for some, video. All dialects have been transcribed orthographically, and some phonetically. Each place is ideally represented with at least one informant of each sex, and sometimes also of different age groups. Part of the corpus is POS tagged, but all of it will be thus tagged in the end. Like with other corpora, an important use is thought to be linguistic searches for words, parts of words or strings of words, and in combination with POS tags. In addition, meta-linguistic variables can be used as search filters.

The corpus is used with the web-based Glossa corpus system (Nygaard et al. 2008), and is user-friendly with pull-down menus and clickable boxes and no need for regular expressions at the interface level. [1]

[1] The Glossa corpus system was developed at the Text Laboratory at the UiO, and is used for a wide range of corpora: monolingual and parallel translation corpora, written and

However, when a corpus includes hundreds of geographical locations from many countries, the typical user will need help to find where the places are actually located. Furthermore, for certain searches, the distribution of hits will in actual fact represent an isogloss for a particular phenomenon. This will not be visible unless the results are projected onto a map.

We have therefore chosen to enhance the corpus with maps. We have chosen to use the Google Maps API in our implementation, given its flexibility w.r.t. projecting information on the maps, their open x and the fact that they cover the whole area. We were looking at other solutions, such as the excellent free online services of the Norwegian Mapping Authority, but since they do not cover the whole Nordic area, they were not an option. in three different ways: 1) for each concordance line, a clickable information button gives geographical information about that hit, 2) the geographical distribution of all the hits are represented in one map, 3) geographical filters can be specified on the map instead of via the list of place names. The first two have been implemented, while the third option is being developed at the moment.

Option 1 gives the name of the place that particular speaker is from (and with age-group, sex, recording year). Importantly, since the place-name may not be known to the researcher, a map is also presented. Figure 1 shows an extract of the search results, with the information button on the left.

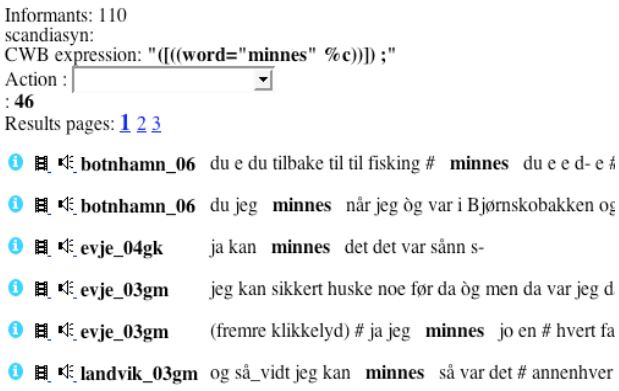

Figure 1: Some hits in the Nordic Dialect Corpus.

By choosing the i-button, the full information about that particular informant (the one from Botnhamn, top row) appears:

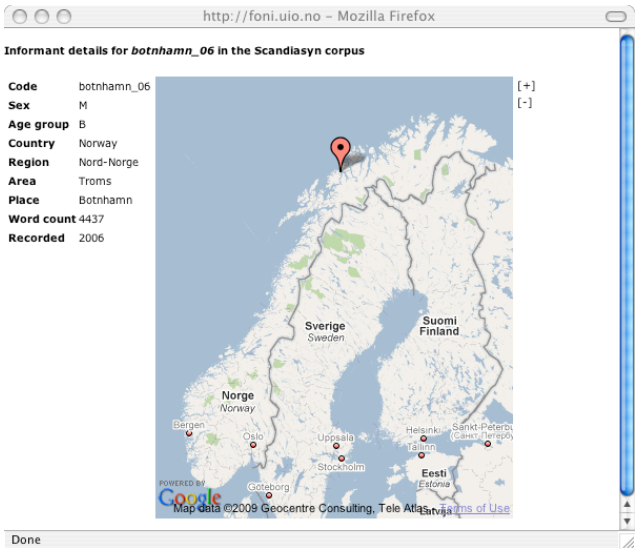

Figure 2: Information on the informant *botnhamn_06*.

This way, the user can quickly determine where a particular example is from. In this case, he is from the island Senja on the coast of North Norway.

Option 2, in a corpus like the Nordic Dialect Corpus, which contains dialects from a geographically widespread area, gives interesting possibilities to see where particular phenomena occur. If the distribution is clearly geographically delimited, this will be an isogloss that is nicely illustrated on the map. For example, we can search for the negation adverb in Norway in the phonetic transcription. (The Norwegian part of the corpus is transcribed both phonetically – following the transcription standard in Papazian and Helleland 2005 – and orthographically.)

We can illustrate by searching in the phonetic transcription for the clitic variants *nte*, *kje* and *kke*; these are all dialectal variants of the equivalent written standards *ikkje* and *ikke* (both meaning 'not'). It is obvious that a long list of hits presented as a concordance would not give the linguist a nice overview of where which form is used, almost no matter how the list is presented or sorted. A distribution indicated on maps, on the other hand, would immediately give us a nice overall picture. We click on a map button that coves the whole resulting concordance.

spoken language corpora. A number of universities use Glossa, due – we think – to its user-friendliness even as more and more search variables are added to it, and its easy access as a freely downloadable system.

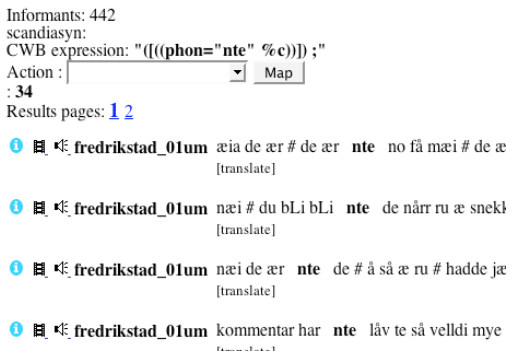

Figure 3: concordance list with map button

Starting with *nte*, we find that this is only used in two places that we have investigated; Fredrikstad and Trysil, both neighbouring Sweden. There are 34 hits in the corpus concordance.

Figure 4: Two places have negator *nte*.

It is not surprising that this is the area where we find it, since *inte/nte* is the main negation word in Swedish.

Turning now to *kje*, we find that this negator covers a big part of Norway (we should add that there are still parts in the white area in the middle that are not yet covered in the corpus).

Figure 5: The negator *kje* is almost everywhere

This result is probably somewhat surprising for many people. The negation *kje* (with its full form *ikkje*) is the one that is used in the written standard *Nynorsk*, which is generally considered to be close to the dialects of West Norway. We see here a much wider geographical distribution. The corpus concordance gives us 3083 occurrences, but it is the map that really shows how widespread it is.

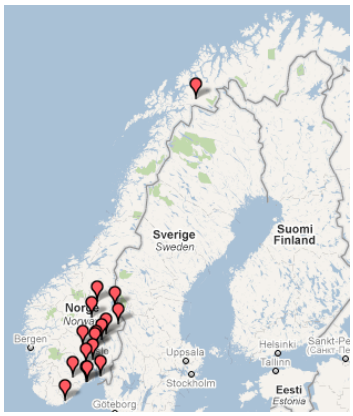

Figure 6: The negator *kke* is quite rare.

The corpus concordance shows 726 occurrences of the negator *kke*. Given that this is the negation (with its full form *ikke*) that is considered the standard in Norway, it is quite interesting that it is rarer both in number and in geographical distribution.

At the moment, only the Norwegian part of the corpus is grammatically tagged. Until we have grammatically tagged all the five languages it will be difficult to search in languages across the whole area, since the orthographies (and lexicons to some extent) differ. We can illustrate a cross-linguistic search, however, by doing a string search: all words starting with *hop*. The result is 318 hits distributed as shown in fig. 7.

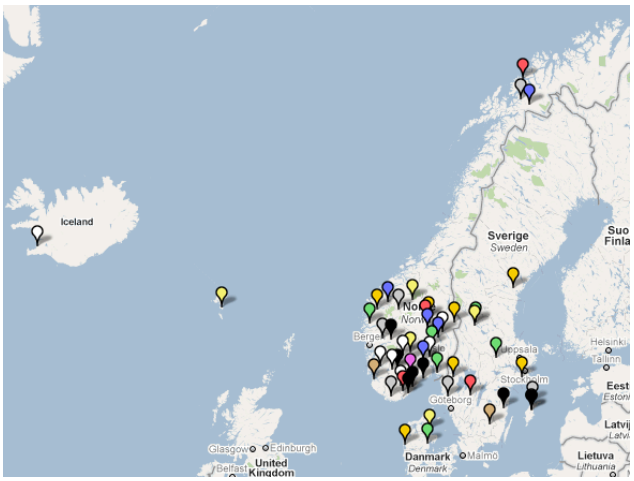

Figure 7: The distribution of the string *hop-*.

This time we got results from (counted from west) Iceland, Faroe Islands, Norway, Denmark and Sweden. Checking the concorance, we find that some actual words are *hoppede* (DK), *hoppa* (NO), *hoppade* (SW), all meaning 'jumped', as well as some country specific ones, like *hoppas* (SW 'hopes') and *hopa* (FA). (The equivalents in the other languages will not be found, due to differences in spelling and lexicon: *håper* (NO), *håber* (DK), *vona* (ICE).)

We think these examples illustrate how useful map illustrations of corpus search results can be.

3. Nordic Syntactic Judgments Database

The Nordic Syntactic Judgments Database is the other research tool developed under the ScanDiaSyn umbrella. It contains speakers' intuitions, i.e. speakers' evaluation of test sentences (many of which are only grammatical in some dialects) presented to them in a questionnaire. The ScanDiaSyn project has gathered data at 270 measure points in Scandinavia, of which 76 places have been put into the database so far. (We believe all the measuring points will be included in 2010.)

A common Nordic pool of around 1400 sentences has been created, and national subprojects have chosen a subset of them. In Norway, 140 sentences are tested, in Denmark 240 sentences. The informant judges each sentence on a scale from 1 (ungrammatical) to 5 (grammatical). Where possible, the database informants are the same ones as those in the corpus, making it possible to test whether what people claim about their language is in accordance with their actual language use. The linguistic literature has shown that their is often divergence between the two kinds of data, and it turns out that our data are no different, as shown with respect to the use of dative case in Norwegian by Johannessen et al. (2009b). In spite of some methodological challenges, judgments questionnaires are indispensable for syntactic research, where some constructions are rare and unlikely to be found in abundance in a corpus, and also because some of the research will be to test which constructions are actually ungrammatical. The maps below will show that the information from the judgments database is actually consistent in each area, and therefore prove their usefulness. More information on the database can be found in Lindstad et al. (2009).

Each test sentence has been appended with a number of linguistic categories describing in as much detail as possible the linguistic property that is tested by that particular sentence. An illustration is given with *wh*-questions differing in the placement of the finite verb, as V3 or V2 (verb in the third or the second sentential position), (1) and (2); with the linguistic description for both given in (3):

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

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

(3) word order, interrog., question, constituent question, simple *wh*-word V3/V2

Querying the database will typically be done in order to find how many have accepted a certain construction and where they are located. The result can be seen as a list of all informants with their judgment for that sentence. We think the database can be used for illustrations at this point, even though at the moment there are only 70 measuring points included. We illustrate with two sentences. One is sentence no. 988 (the same as (1) above), and one is an exclamation, sentence no. 311, here (4):

(4) Hva biler det var her
what cars it was here
'What a lot of cars there are here!'

The result is presented as in figure 10 at the end of this paper. The judgments are represented both by number and colour, where red colour and low number mark means that a particular sentence is considered ungrammatical by a particular speaker in this particular dialect, while green colour and high number mark the opposite.

In our map application, we can choose a number of options. We illustrate with the top-most choice in figure 8. Here we get, for the sentences or categories we have chosen, all the locations where the informants have valued the sentence at 4-5 on average, i.e. accepted the sentence as fully grammatical. Seeing a long list of place names with individual judgments, like in figure 10 at the end of this paper, would not be as revealing as a map.