A corpus of spoken Faroese

Janne Bondi Johannessen

University of Oslo

Abstract:

The paper describes the new Corpus of Spoken Faroese. While the corpus is still under development with respect to content (number of informants, dialects and words), it is included in the larger Nordic Dialect Corpus, which means that all technical solutions are already in place. As a result, the Faroese corpus is fully operable, albeit with a rather limited number of words at present. The recordings have all been made, but transcription and tagging remain undone for most of them, however these are expected to be finished by the end of 2009. At the moment, there are nine conversations in the corpus. In the paper I describe some of the search and resrult-handling options the corpus offers, exemplifying with Faroese, and I also try to shed light on some linguistic questions using the corpus.

1. Introduction

The Corpus of Spoken Faroese consists of recordings done in the summer of 2008.[1] The corpus is still under development as this paper is being written. However, what is ready is enough to give potential future users a good idea of what it is like and how it can be used. In section 2, I will give a description of the bigger picture of which this corpus is a part. I will then give an overview of the contents of the corpus, and give a thorough presentation of search possibilities and results handling using the corpus interface. In section 3, I illustrate how the corpus can be used to resolve some linguistic issues. In this section it will also be clear that the corpus has a better potential for full use once all the recordings have been included and the texts grammatically tagged.

---

[1] The recordings were done in connection with the 5th NORMS Dialect Workshop in the Faroe Islands. They were performed by Janne Bondi Johannessen and assistant Karine Stjernholm from the Text Laboratory, University of Oslo, with local recording assistants Rakul Napoleonsdóttir Joensen and Petra Eliasen. Eliasen has also transcribed the recordings orthographically. Jógvan í Lon Jacobsen and Victoria Absalonsen were the local organisers who made all the necessary arrangements, and Zakaris Svabo Hansen was very helpful towards the practical questions concerning recording locations. Eliasen has kindly helped translate the corpus examples used in this paper. I am grateful to Caroline Heycock for her good comments and suggestions. The recordings were financed by NORMS (i.e., NOS-HS) and the transcriptions and technical development by NordForsk and the UiO.

Nordlyd 36.2: *NORMS Papers on Faroese*, 25–35. Edited by Peter Svenonius, Kristine Bentzen, Caroline Heycock, Jógvan í Lon Jacobsen, Janne Bondi Johannessen, Jeffrey K. Parrott, Tania E. Strahan, and Øystein Alexander Vangsnes
CASTL, Tromsø. http://www.ub.uit.no/munin/nordlyd

2. Description of the corpus

2.1. Background

Language research, and especially dialect research, depends on empirical data. For many researchers through the history of dialect studies, this has meant that the researcher has to go to the relevant area and collect his or her own material. Clearly, this is a very time consuming task. Further, it also means that much work has had to be duplicated, since recordings have not had an easy way of being shared. Written transcriptions sometimes have been used, but since transcription is very expensive to do, not many have been able to make this extra effort.

With the invention of electronic corpora that can be distributed via the internet, empirical language studies have become cheap and easy. They represent in some sense a democratisation of research possibilities. Corpora containing speech are very expensive compared with written language corpora because their development is labour-intensive, all the way from collection to transcription. Making available audio (and video) requires much storage space and specialised servers, adding to the cost. It is therefore a very good use of resources once a corpus has been created, to share it with other researchers. The Corpus of Spoken Faroese is open to all researchers.

An important bonus of electronic corpora is the way large amounts of data can be studied quickly. The researcher does not have to read through each word in each text to find out whether a given phenomenon exists in the material. Advanced search options make it possible to go through large collections in very little time. This method has at least two fortunate types of results: First, the reseacher can get lots of examples that will shed light on a particular research question, and second, the researcher may get examples that are not amongst those originally studied, and can serve to increase the understanding of a given phenomenon, or even point the researcher to a new and previously unstudied one.

The corpus is part of the Nordic Dialect Corpus, a common effort under the umbrella of the Nordic research network ScanDiaSyn and the Nordic Centre of Excellence in Microcomparative Syntax. The technical development is being done by the Text Laboratory, University of Oslo, while the individual national recordings are usually the responsibility of the individual countries. Thus, the Swedish spoken data have been provided by the project Swedia 2000, with transcriptions by the project SweDiaSyn. The Danish recordings by DanDiaSyn, and the Norwegian ones by NorDiaSyn. From Iceland recordings are still in the making, while for Faroese, a joint effort has been necessary. Thus, these recordings have been performed as a collaboration between the University of Oslo (The Text

Laboratory) and the local Faroese organisers of the 5th NORMS Dialect Workshop.

2.2. Corpus contents

The corpus will consist of recordings of young (under the age of 30) and old (over 50) speakers, both women and men, from five places: Tórshavn, Fuglafjørður (Eysturoy), Klaksvík (Borðoy), Tvøroyri (Suðuroy), and Sandur (Sandoy). Each speaker (informant) takes part in a ten minute interview with a recording assistant, and an informal 30 minute conversation with another local speaker. No informant should speak about private and confidential matters or about friends and family, so the informants were presented with a list of topics to facilitate the dialogue. These topics were carefully chosen to inspire the Faroese informants, and included things like a controversial author, sheep farming, and road tunnel development.

At present, the corpus consists of nine informants, but the final corpus will have 20 informants. The number of words is nearly 22 000, but this will increase to an expected 150 000 words, since most of the long conversations have not been included yet. The corpus will be grammatically tagged with parts of speech, which will enable generalised searches.

2.3. Corpus interface

The Corpus of Spoken Faroese is part of the bigger Nordic Dialect Corpus. This corpus is built with the Glossa search system whisch is built on top of the Corpus Workbench CQP system. This ensures a very user-friendly system, which is at the same time advanced w.r.t. search options and results handling.

2.3.1. Searching the corpus

Since the corpus is part of the Nordic Dialect Corpus, all languages can be searched for at the same time, unless the user chooses a particular language.

Figure 1. How to get from Nordic Dialect Corpus to Corpus of Spoken Faroese.

The interface gives ample opportunities for simple and advanced search. For example, it is possible to search for words that start or end with a particular letter sequence, or to exclude a search word, or combine sequences of words. These linguistic searches can be combined with filtering the informants according to age, place or gender. We will illustrate some of these options below.

In order to search for words beginning with the sequence *be-* the user can simply choose the alternative filtering from a neat menu below the typing box, in which s/he can write *be*.

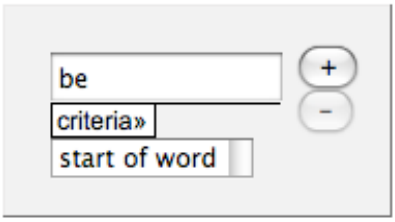

Figure 2. How to search for words beginning with be-.

If this system of typing and pull-down menus did not exist, the user would have to formulate instead a regular expression – the only option in many corpus systems:

(1) CWB expression: "([((word="be.*" %c))]) ;"

The results, at present 79 hits, are presented like a concordance:

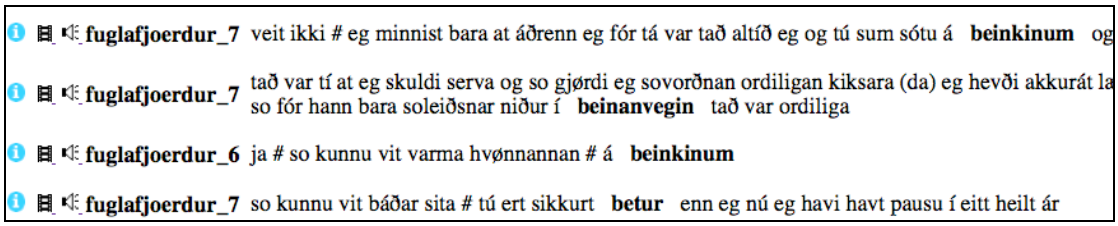

Figure 3. Some of the search results for be-.

We will see more about search results and ways of handling them in the next subsection.

It is also possible to search for more than one word in combination. Below is illustrated a search for words beginning with *be-* followed by the word *og* with at most 3 words in between.

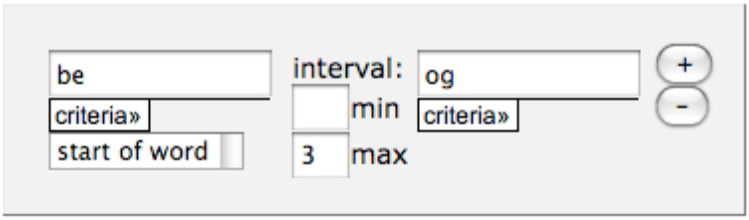

Figure 4. Searching for a word starting with be- followed by og with at most 3 words in between.

Again, an alternative search using regular expressions is somewhat more complicated:

(2) CWB expression: "([((word="be.*" %c))] []{0,3} [((word="og" %c))]) ;"

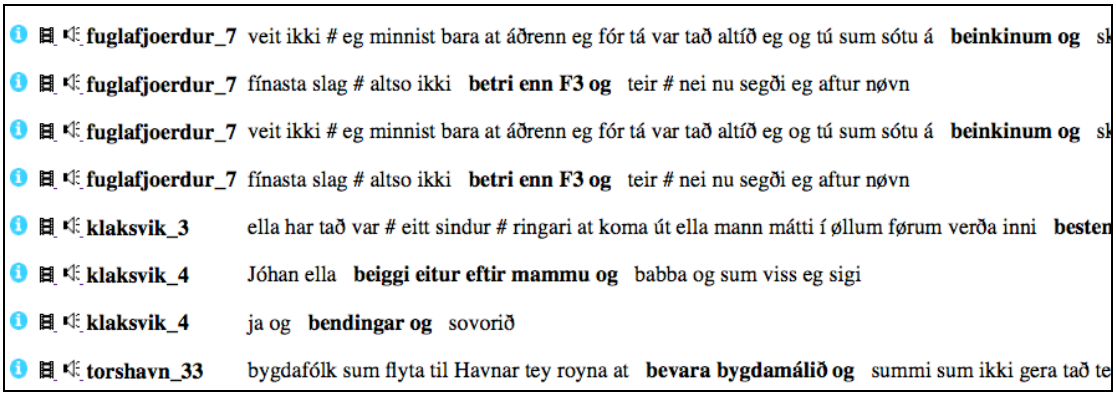

Figure 5. Results for the search presented in Figure 1.

If the investigator wishes to filter the linguistic search through social variables such as age and gender, this is done by clicking on the expandable buttons Age and Sex, and then choose the desired options:

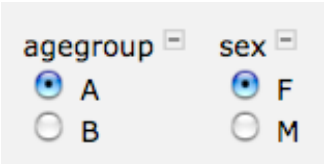

Figure 6. A search enriched with filters for age (young) and sex (female).

To check that the program has chosen the correct informants for this search, we can click on a button Show Texts, which gives this information:

Code	Sex	Age group	Country	Region	Area	Place	Word count
fuglafjoerdur_6	F	A	Faroe			Fuglafjørdur	1,638
klaksvik_3	F	A	Faroe			Klaksvik	3,908
fuglafjoerdur_7	F	A	Faroe			Fuglafjørdur	3,849
klaksvik_4	F	A	Faroe			Klaksvik	3,561

Figure 7. The Show Texts button gives an overview of the informants chosen for a particular search. (Here: young females from the Faroe Islands.)

2.3.2. Results handling

In figure 5 we saw the standard search results presentation. Notice that there are some symbols to the left of each line. The i-symbol can be clicked on and then shows information about that particular informant. For example:

Informant details for *fuglafjoerdur_6* in the Scandiasyn corpus

Code	Sex	Age group	Country	Region	Area	Place	Word count	Recorded
fuglafjoerdur_6	F	A	Faroe			Fuglafjørdur	1638	2008

Figure 8. Information given in the clickable information box.

The film icon is more exciting. Clicking on it gives the video clip, with audio, that a particular transcription is based on.

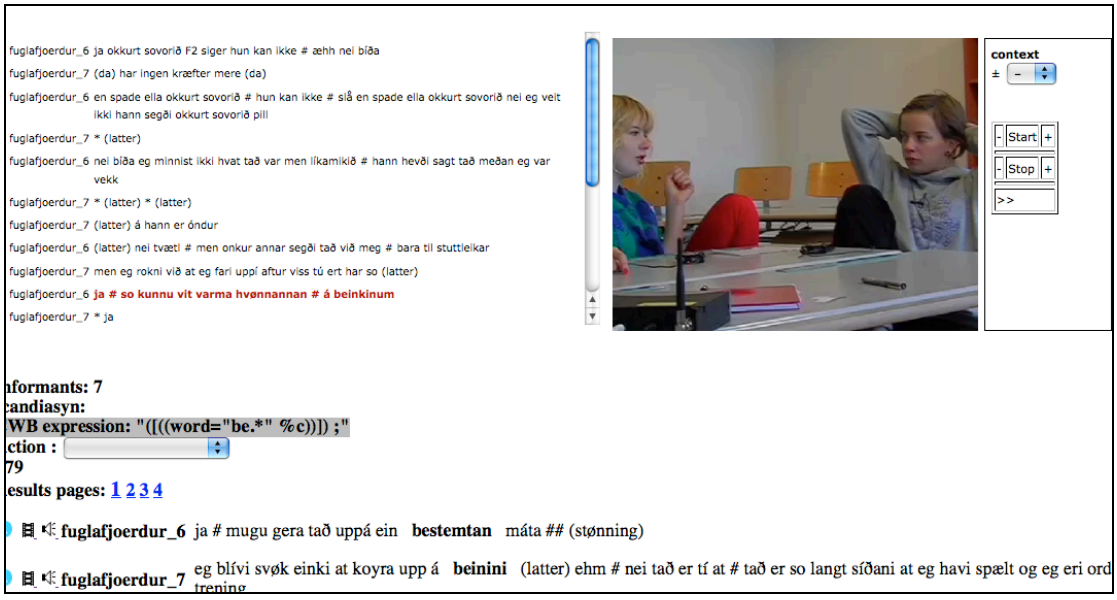

Figure 9. Video and audio accompanying transcription.

On the results page is also an action menu. By choosing Count from that menu, the words that satisfy the search are presented in a frequency sorted list, or a pie chart (other options are also available):

occurences	match
12	betri
10	beint
5	beinanvegin
4	betur
4	beinkinum
4	bestemt
3	bevíst
3	ber
2	bestemtan
2	bensin
2	bera
2	bestemmað

Figure 10. Results as frequency list.

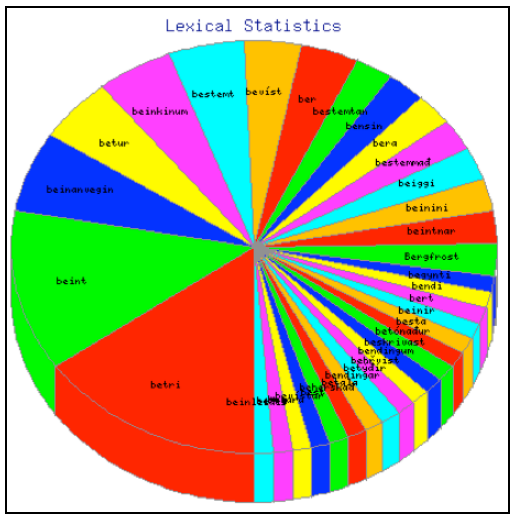

Figure 11. Results as pie chart.

It is of course possible to save the search results. But it is also possible to further process the results. The action menu gives the opportunity of saving or deleting hits. If these options are chosen each concordance line is presented with a box on the left, which can be ticked for further treatment of particular hits. Below is an illustration of the results with the delete option. One hit is is ticked; that of a proper name, which would possibly not be interesting in a linguistic search.

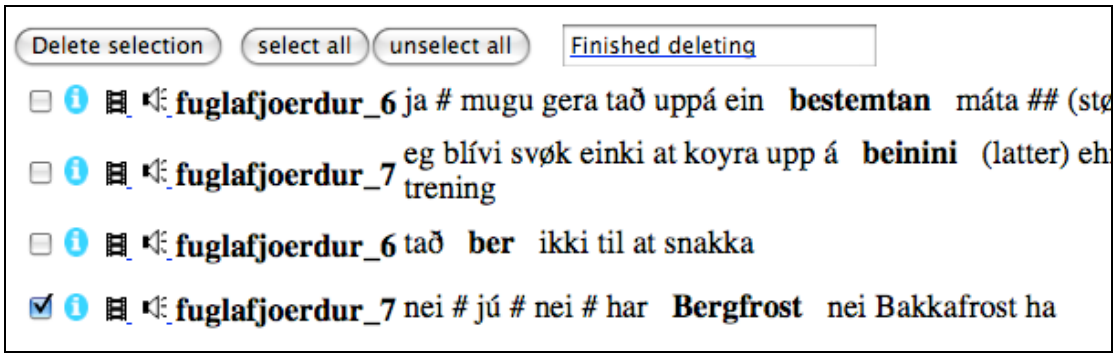

Figure 12. Example of the delete option in the action menu.

It is also possible to annotate the hits. For example, the user could choose to create a tag set consisting of PN (proper names) and CN (Common nouns), and the annotate option would then give the possibility of annotating each concordance line with one of those tags.

3. Attempting to answer linguistic questions using the corpus

Thráinsson and Hansen (2008) summarise a number of claims regarding Faroese syntax. In this section I will see whether I can shed light on some of them by using the corpus, even in its present, unfinished and small state.

3.1. The position of adverbs in subordinate clauses

Let us start by looking at the position of adverbs in subordinate clauses. Thráinsson and Hansen (2008) refer to Barnes (2001), Vikner (1995), Thráinsson, Petersen, Jacobsen and Hansen (243ff.), and Petersen and Adams (2008:281), and cite the following patterns of judgments:[2]

(3) Fa: Hann spurdi, hví Jógvan ?**las ikki / ikki las** bókina.
Ic: Hann spurði, af hverju Jón **læsi ekki / ?*ekki læsi** bókina.
Da: Han spurgte, hvorfor Jens ***læste ikke / ikke læste** bogen
he asked why Jens read not / not read book.the

Thus, they state that Faroese is more like Danish (and Norwegian and Swedish) in allowing negation to precede the finite verb in subordinate clauses, while the other order is intermediate in status. To test this, it would be very useful if the corpus were tagged, as it will be in the not too distant future. Then we could simply search for a subjunction followed by an adverb with an appropriate number of words in between. As it is now, we have to specify individual words, like *sum* ('which') and *ikki* ('not'):

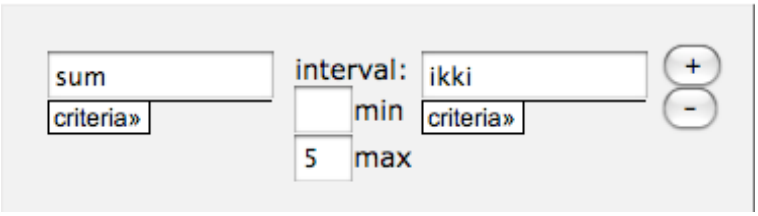

Figure 13. Trying to find the relative position of adverb and verb in subordinate clauses, by specifying each word class by individual words.

This is not ideal, since each such search gives very few hits. In the present, admittedly small, corpus the result was 13 hits. However, the tendency is clear. Although some hits were irrelevant because the adverb was in a different clause than the subjunction, all the relevant hits have the order adverb–finite verb, and thus confirm the pattern suggested at in (3); that this order is the preferred one in subordinate clauses in Faroese, just like in Danish. Two examples are given below:

[2] This topic is also the theme he theme of Angantýsson (2009) and one of the other papers in this volume of another paper in this volume, by Kristine Bentzen, Piotr Garbacz, Caroline Heycock, and Gunnar Hrafn Hrafnbjargarson.

(4) a. hugsa um landssjúkrahúsið sum næstan ikki hevur ráð
remember about hospital that almost not has afford
til medisin
to medicine
'Remember the hospital, which had almost no money for medicine.'
b. og tað sum ikki hevði við skúlan at gera
and that which not had with school.the to do
'... and that which did not have to do with the school.'

3.2. Null expletives

Thráinsson and Hansen (2008) also cite Thráinsson, Petersen, Jacobsen and Hansen (284ff) in saying that null expletives are ok in Farose:[3]

(5) Fa: Nú regnar (**tað**) heldur illa.
Ic: Nú rignir (***það**) mjög mikið.
Da: Nu regner *(**det**) helt vildt.
now rains (it) completely wildly

Unfortunately, with the verb *regnar* ('rains'), there is only one hit:

(6) tað bara regnar i dag
it just rains in day
'It just rains today.'

This is irrelevant, since, for independent reasons, it is expected that something has to fill the clause-initial position. Thus, relevant examples would have had to have an adverb clause-initially. At present, then, the corpus cannot give us any indication in either direction for this syntactic feature, at least with this combination of search words. When the corpus is tagged, it will be possible to make more generalised search expressions, by for example searching for the string adverb plus finite verb.

3.3. Possible adverb between subjunction and subject

Thráinsson and Hansen (2008), citing Barnes (1992), also point out a difference between Faroese and Swedish (and Norwegian); there can never be an adverbial between the subjunction and subject or whatever else sits in the next position.

[3] See also Angantýsson (2009).

(7) Fa: *Tað er synd, at **ongantíð** Jógvan kann koma.
Ic: *Það er slæmt, að **aldrei** Jóhann getur komið.
Sw: Det är hemskt, att **aldrig** Johan kan komma.
it is sad that never Jogvan can come

Querying the corpus for the subjunction *at* 'at' gives very few cases of this pheonomenon, if any. There are 588 occurrences of *at*, but many, possibly most, are examples of the homophonous infinitival marker. There is, actually, one occurrence in which there is a possible analysis in which the adverb *ikki* 'not' follows a subjunction:

(8) men eg haldi at ikki at, nú veit ikki akkurát hvussu
but I think that not that, now know not exactly how
stevnurnar vóru í trýssunum og hálvfjersunum
festivals were un the-sixties and seventies.the
'But I don't think that, that, now (I) don't know exactly how the festivals were in the sixties and seventies.'

But since the sentence is interrupted, we clearly cannot be sure. No other example occurs amongst the hits. For this particular example a bigger corpus is necessary in order to state something about the general tendency. It would also be an advantage if the corpus were tagged, so that more general searches could be made, for example by searching for subjunctions in general, without getting false hits like the infinitival markers, and also for being able to specify the sequence subjunction plus adverb.

4. Conclusion

In this paper I have presented The Corpus of Spoken Faroese, which is a subpart of the bigger Nordic Dialect Corpus. The Corpus of Spoken Faroese is under development, and many recordings still await transcription and grammatical annotation. We have seen some reasons for using an electronic corpus. We have also looked at the search interface and results handling system, with examples to show how the corpus can be used. We have also tried to look at some linguistic issues that the corpus could shed light on. For some phenomena it turns out to be possible to use the corpus even in its infant stage, but for other phenomena the corpus will be more useful once it is bigger and grammatically annotated.

References

Angantýsson, Ásgrímur. 2009. Verb/adverb placement and fronting in embedded clauses in Faroese. Ms. University of Iceland.

Barnes, Michael P. 1992. Faroese Syntax — Achievements, goals and problems. In Jonna Louis-Jensen and Jóhan H.W. Poulsen (eds.), *The Nordic Languages and Modern Linguistics* 7, pp. 17–37. Føroya Fróðskaparfelag, Tórshavn. [Also in *Scripta Islandica* 43:28–43 and in Barnes 2001.]

Barnes, Michael P. 2001. *Faroese Language Studies*. Studia Nordica 5. Novus, Oslo.

Bentzen, Kristine, Piotr Garbacz, Caroline Heycock, and Gunnar Hrafn Hrafnbjargarson. This volume. On variation in Faroese verb placement.

Petersen, Hjalmar P. and Jonathan Adams. 2008. *Faroese. A Language Course for Beginners*. Sprotin, Tórshavn.

Thráinsson, Höskuldur, Hjalmar P. Petersen, Jógvan í Lon Jacobsen, and Zakaris Svabo Hansen. 2004. *Faroese. An Overview and Reference Grammar*. Føroya Fróðskaparfelag, Tórshavn.

Thráinsson, Höskuldur and Zakaris Svabo Hansen. 2008. A Comparative Overview of Faroese Syntax. Paper presented at the 5th NORMS Dialect Workshop, Tórshavn.

Vikner, Sten. 1995. *Verb Movement and Expletive Subjects in the Germanic Languages*. Oxford University Press, Oxford.

Web sites

DanDiaSyn: http://www.hum.au.dk/dandiasyn/
Glossa: http://www.hf.uio.no/tekstlab/English/glossa.html
NorDSiaSyn: http://www.tekstlab.uio.no/nota/scandiasyn/nordiasyn.html
Nordic Dialect Corpus: http://www.tekstlab.uio.no/nota/scandiasyn/index.html
NORMS: http://norms.uit.no/
ScanDiaSyn: http://uit.no/scandiasyn
Swedia 2000: http://swedia.ling.gu.se/
SweDiaSyn: http://uit.no/scandiasyn/swediasyn/
Text Laboratory, UiO: http://www.hf.uio.no/tekstlab/
5th NORMS Dialect Workshop:
http://norms.uit.no/index.php?page=foroyar_program