Generating a lexicon of Scandinavian modals

Gunnar Hrafn Hrafnbjargarson
The Text Laboratory, University of Oslo

Abstract:

Morphology and phonology can in many cases be used to figure out which words correspond to which in Scandinavian. For instance, it is rather easy to figure out which Norwegian personal pronoun corresponds to which in Danish, and even Icelandic or Faroese. However, when it comes to prepositions and modal verbs we cannot rely on morphology or phonology alone. For example, Norwegian and Danish *måtte* do not always have the same meaning, and similarly, Icelandic *vilja* is not used as the future modal as Norwegian and Danish *ville* is. Instead of relying on morphology or phonology, we can use parallel corpora. Unfortunately, there are not many parallel corpora that include **all** of the Scandinavian languages, and those that exist are maybe not large enough to give reliable results. Nevertheless, to get a picture of what it could look like, the Danish, Faroese, Icelandic, Norwegian, Swedish, and English parts of a small treebank, *The Sophie Treebank*, were used to find out which modal verbs correspond to which in the various languages.

1. ScanLex[1]

1.1 An inter-Scandinavian lexicon

One of the goals of the Scandinavian Dialect Syntax project is to establish a database with dialect material from all of the Scandinavian (North Germanic) language varieties, i.e. Danish, Faroese, Icelandic, Norwegian, and Swedish. The database will be accessible through an online search interface.

Each of the Scandinavian varieties will be represented in the database with transcriptions of interviews or discussions between two or more speakers of the same dialect, or written elicitation task reports, questionnaires, etc. One type of transcription will be a transcription where the dialect is "translated" into a written standard. For example, Swedish dialects will be transcribed in standard Swedish, and Norwegian dialects

[1] The research reported in this article was financed through a grant from Nordens Sprogråd (Nordic Language Council / Nordic Council of Ministers), project no. 254440-05127. I would like to thank Janne Bondi Johannessen and Chris Biemann, in particular, for discussions and help and, not least, Kristin M. Eide for kindly providing me with literature on and guiding me through the jungle of Norwegian modals. Usual disclaimers apply.

Nordlyd 34: 183-199, © Hrafnbjargarson 2007
Scandinavian Dialect Syntax 2005
Edited by Kristine Bentzen and Øystein Alexander Vangsnes
CASTL, Tromsø. http://www.ub.uit.no/munin/nordlyd

will be transcribed in one of the two Norwegian written standards, *bokmål* or *nynorsk*.[2]

The transcriptions will finally be tagged with morphological (and sometimes syntactic) information, which enables not only search for specific lexical items or parts of words, but also syntactic patterns. With respect to search for specific lexical items, we have a potential problem that we have to solve unless we do not process the data in the database further. Firstly, people with limited or no knowledge of Scandinavian will practically be excluded from using the application. Secondly, even people with sufficient knowledge of Scandinavian will run into potential problems because very few people know the exact equivalents of some word in the other Scandinavian varieties.

One solution to this problem is to lemmatize all of the texts in all of the written standards, plus English, or in more general terms, to apply every word in a transcription with a lexical entry from all the other languages. For this type of multi-language lemmatization, we need a (inter-Scandinavian) lexicon, with 'one-to-one' correspondences.[3] The example below shows how the Icelandic pronoun *hún* 'she.NOM' can be represented in the lexicon:

(1)

Ic.	Nob.	Da.	Fa.	Non.	Sw.	En.
hún	hun	hun	hon	ho	hon	she
	ho					

From the example in (1), we see that, in theory, it can be possible to generate a lexicon of this sort by only looking at morphological or phonological features. But as soon as we want to generate the lexicon (semi-) automatically and as soon as we look at other parts of the pronominal system (not to mention outside the pronominal system) we run into problems. For example, the accusative form of the feminine singular pronoun in Icelandic, *hana*, is phonologically/morphologically closer to the Norwegian masculine singular pronoun *han* than to *ho* (which can also be the object form). Therefore we need large parallel corpora; either corpora

[2] Norwegian has two written standards *bokmål* and *nynorsk*, sometimes called Dano-Norwegian (because of its relation to Danish) and Neo-Norwegian in English. Throughout the paper, I will use the Norwegian terms for the two standards abbreviating *bokmål* as *Nob* and *nynorsk* as *Non*.

[3] As can be seen in (1), the phrase *'one-to-one' correspondence* does not actually mean that each word only has one and only one correspondent in the lexicon. What it means is that each word will be represented with at least one correspondent in the lexicon.

where all of the Scandinavian languages are represented at the same time, or corpora where one or more Scandinavian languages are parallelled with English (which then can serve as a link between the other languages that are not represented in the respective corpus). However, very few such corpora exist for Icelandic and Faroese, and in addition, we also have to develop methods for extracting data from large corpora.

1.2 The database

ScanLex is a mySQL database that is made of three tables:

(2) Tables in Scanlex
1 Lemma and language
2 Form, morphological tags, and additional information
3 Sense (semantic classes) and (semantic) descriptions

The three tables are joined (linked to each other) through an identification number given to each entry in the first table and (for pronouns) through the different word forms.

The first table consists of three columns: `lemmaid` where each dictionary entry (or lemma) is given an identification number, `lemma` for the dictionary entry itself, and `language` where it is specified which language the word comes from.

Table 1 is linked to Tables 2 and 3 in such a way that `lemmaid` is repeated throughout the tables.

(3) Table 1. Lemma and language

lemmaid	lemma	language
1	ég	Icelandic
2	þú	Icelandic
3	hann	Icelandic
4	hún	Icelandic
5	það	Icelandic
...		
217	skulle	Swedish
218	vilja	Swedish
219	must	English
...		

Each lemma (or dictionary entry) gets **one row** in the table.

Each entry is specified for which **language** it comes from.

Each entry gets a **lemma identification number.**

The second table consists of nine columns. As in Table 1, each row gets an identification number (`id`), but here, the id is not important for the

other tables, it is only there to reduce risks of making mistakes when editing the database.

(4) Table 2. Forms and morphology

id	form	person	lemma_id	number	kasus	gender	url	extra
1	ég	1	1	sg	nom		is.html	
2	mig	1	1	sg	acc		is.html	
3	mér	1	1	sg	dat		is.html	
4	mín	1	1	sg	gen		is.html	
...								
52	hansara	3	13	sg	gen	masc	fo.html	
53	hon	3	14	sg	nom	fem	fo.html	
...								
168	dem	3	38	pl	obj		nob.html	
...								
1213	skula		202				fo.html	modal verb
1214	skulle		209				dk.html	modal verb
1215	kunne		210				dk.html	modal verb
1216	ville		211				dk.html	modal verb
...								

Every form gets an **entry.**

Pronouns are specified for **person, number, case, and gender**.

Each entry is related to **lemma** through the lemma identification number.

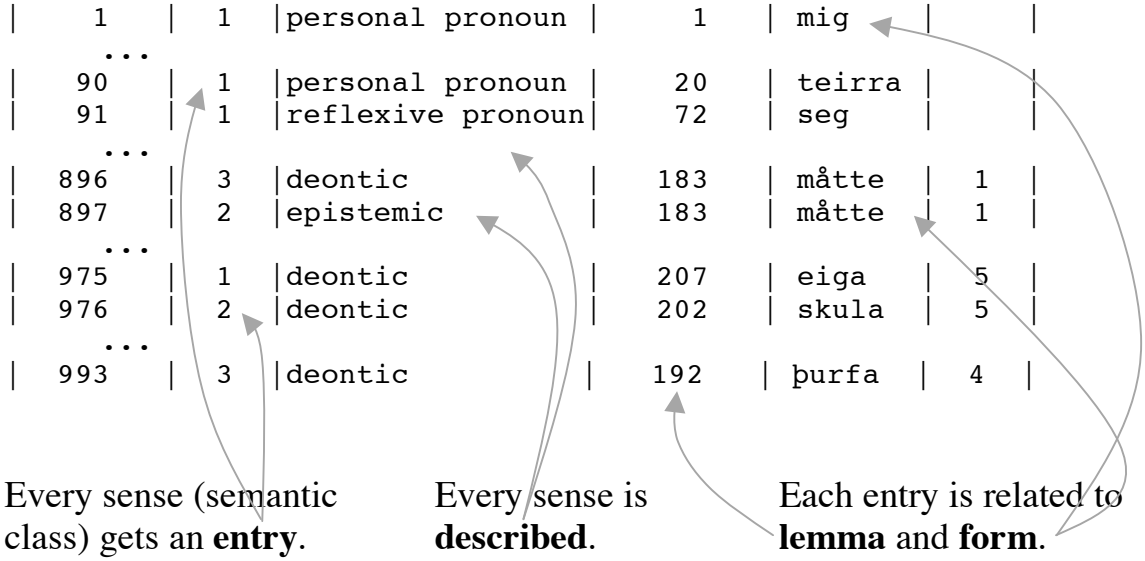
As in Table 1, each `form` (instead of each dictionary entry) gets one row. Each form is specified for `person`, `number`, case (`kasus`), and `gender`. Thus, the form *mér* in Table 2 is first person singular dative. The modal verbs that are shown in Table 2 are not specified for person or number since the forms that are shown are the infinitival forms. When using the search interface, see below, the information in the column `url` is used to generate links to a page where the declination/inflection of the word is shown. The column `extra` is used for additional information from reference grammars if for example different terminology has been used (*determiner* vs. *demonstrative*) or whenever reference grammars refer to a specific use of a certain word (e.g. whether the word is rare, informal, whether the word mostly occurs in the spoken language or predominantly in the written language). Finally, `extra` is used for redundant information such as type of verb, e.g. *modal verb* or *auxiliary verb*.

Table 3 consists of six columns. The numbers in the column `sense_id` are used as identification numbers when editing the table. In the column `sense`, different semantic meanings of each form are listed and in `description`, these semantic meanings are described in a few words. Each sense is related to a lemma in Table 1 in the column `lemma_id2` and a form in Table 3 through the column `form_2`. The last

column, `type`, was needed to distinguish different types of epistemic/deontic modals from each other.

(5) Table 3. Sense and semantic descriptions

sense_id	sense	description	lemma_id2	form_2	type
1	1	personal pronoun	1	mig	
...					
90	1	personal pronoun	20	teirra	
91	1	reflexive pronoun	72	seg	
...					
896	3	deontic	183	måtte	1
897	2	epistemic	183	måtte	1
...					
975	1	deontic	207	eiga	5
976	2	deontic	202	skula	5
...					
993	3	deontic	192	þurfa	4



Although the primary use of ScanLex will be to enable the multi-language lemmatization of the ScanDiaSyn database, the database structure of ScanLex makes it possible to access the lexicon through an online search interface. The interface will have several possibilities, such as search for specific lexical items in different languages (where the user can define in which language he/she wants to know the corresponding word) with or without additional morphological information, or to search for words that belong to specific word classes, e.g. indefinite pronouns, prepositions that govern dative, etc. At the moment a simple provisional search interface can be accessed at http://omilia.uio.no/scanlex/.

2. Using a parallel corpus for extracting modal verb correspondences

2.1 The corpus dilemma

As I have mentioned, it is possible to use morphological and phonological similarities to find correspondences in some parts of the lexical inventory of Scandinavian. For instance, it is relatively easy to find correspondences within the pronominal system and so far, pronouns have been fed manually into ScanLex (Norwegian *han* 'he.NOM' = Icelandic *hann* 'he.NOM') mostly through morphology, i.e. person, number, gender, case, but sometimes also meaning. However, phonology and morphology are far from being sufficient decisive factors when we move outside the pronominal system. Although the Norwegian modal *skulle* looks (and almost sounds) exactly

like Danish *skulle*, these two do not always have the same meaning. Likewise, the Danish modal *ville* does not have an Icelandic correspondent *vilja* (unless in the root sense). Since we cannot rely on morphology or phonology here, we have to find something else, for instance parallel corpora.

The obvious problem is, as I have mentioned, that not many parallel corpora cover all of the Scandinavian languages. One such corpus is *The Sophie Treebank*[4], which I have used here; another one is the *KDE* part of the *OPUS corpus*[5] although this part of the *OPUS corpus* does not include Faroese.

2.2 The Sophie Treebank

The Sophie Treebank is a parallel treebank that consists of material from ten 'North European' languages: Danish, Dutch, English, Estonian, Faroese, Finnish, German, Icelandic, Norwegian, and Swedish. The text is taken from the Norwegian original and the translations of the two first chapters of Jostein Gaarder's novel *Sofies verden* (*Sophie's World*). In my survey, I used the Scandinavian parts (Danish, Faroese, Icelandic, Norwegian, and Swedish) in addition to the English one. The texts contain approx. 530-550 sentences, except the Swedish one, which only contains the first chapter (215 sentences).

2.3 The method(s)

I started by collecting all sentences containing a modal from each of the languages, one by one. The number of sentences that contain a modal in the corpus is shown in (6), where modal means a verb that takes a bare verb as its complement (although I also included the Icelandic and Faroese modals that require an infinitive marker).

There were 171 sentence "pairs" or aligned "blocks." In most cases, a sentence with a modal in some language had a correspondent in some other language, but in some cases, there were no correspondents. The numbers following the forward slash in (6) indicate the total number of sentences for each language:

[4] Go to http://hf.uio.no/tekstlab/prosjekter/SOFIE.htm for more information on *The Sophie Treebank*.

[5] The KDE part contains KDE system messages (KDE = *K Desktop Environment*, a user interface for Linux and UNIX systems). Go to http://logos.uio.no/opus/ for the *OPUS corpus*.

(6) Sentences with modals in *Sofies verden* / Total:

Danish: 126 / 547

Faroese: 85 / 556

Icelandic: 101 / 542

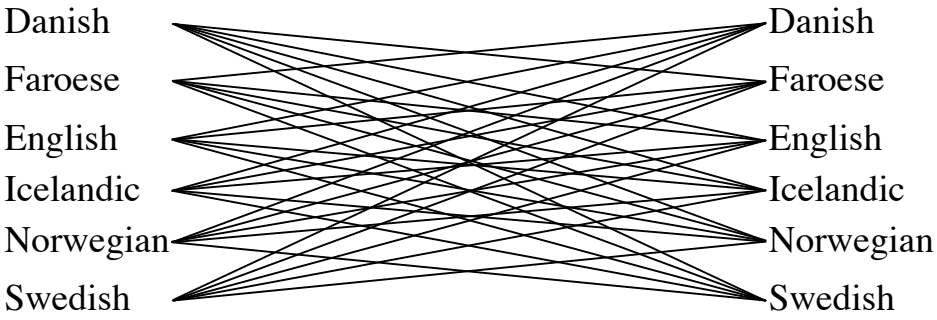
Norwegian: 124 / 543

Swedish: 48 / 215

English: 84 / 530

In the beginning, I wanted to use the material that I extracted from the corpus to find all possible matches from each language to the other languages in a process that can be described as follows:

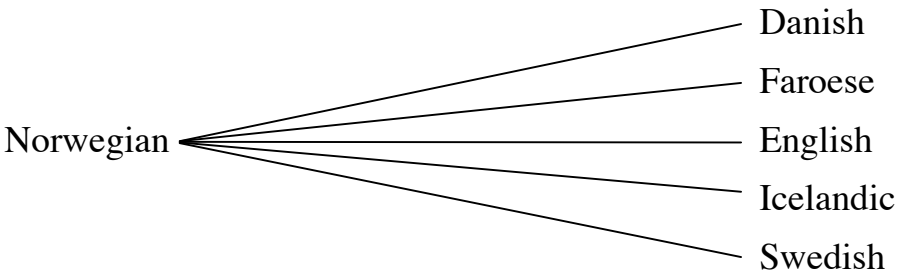
(7)



With this method, every sentence with a modal in each language was linked to a sentence in the other languages. This method turned out useless, mainly because the results were not only correspondences between modals in the different texts, but also correspondences between modals and other types of verbs. The results were different heterogeneous lists of correspondences for each starting point.

The alternative method turned out to be more fruitful. First, I grouped the modals in the Norwegian original roughly according to the definitions of the Norwegian Reference Grammar (*Norsk referansegrammatikk*, Faarlund et al. 1997:579-637), i.e. epistemic modality vs. deontic modality. Then, I found the corresponding modals in the five translations:

(8)

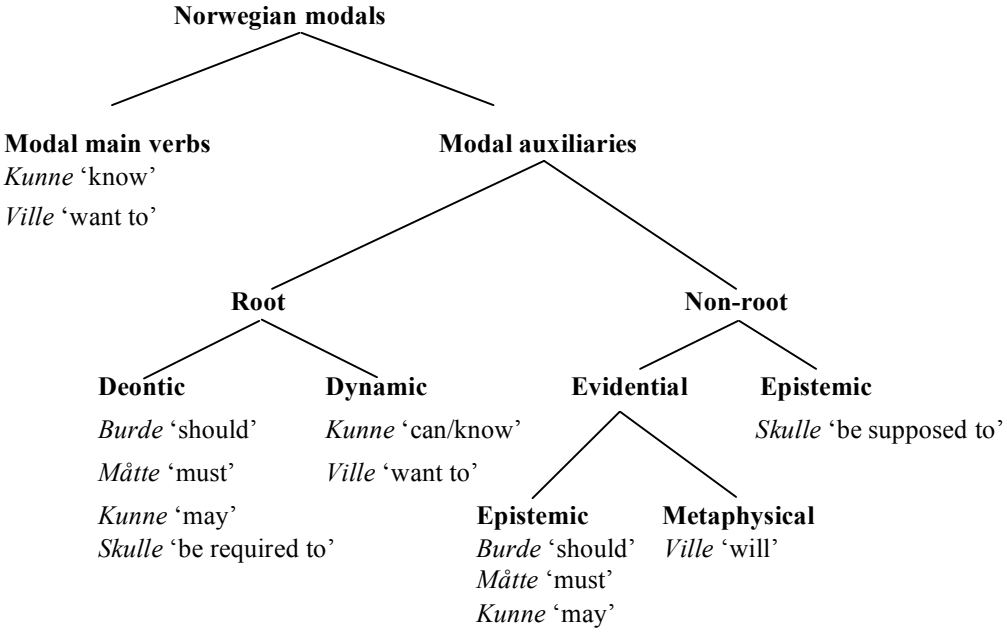


Since the grouping of the modals was made very roughly, the database still needs to be "proof-read," i.e. to make a more fine-grained analysis and

preferrably, such an analysis should be made according to the analysis presented in Eide (2005).

Eide's division of Scandinavian modals is very detailed. She shows, for instance, that Norwegian modals should be divided into six different groups, shown in the figure below, according to both syntactic and semantic factors.[6]

(9) Norwegian modals (adapted from Eide 2005)



Although my division is very sketchy, the preliminary results can in fact be viewed independently of whether the respective modals show epistemic or deontic modality, because the results can be used to see which modals are used to translate a specific Norwegian modal in Danish, Faroese, Icelandic, Swedish, and English. Since the corpus is very small, the results should not be considered final or conclusive in any way.

At the moment we only have resources for partial automatic analysis of such small data, i.e. it is possible to automatically extract the data and align it (on sentence, or even word basis, cf. Tiedemann 2003) but we do not have resources for automatic semantic analysis, i.e. to distinguish epistemic and deontic modality. Therefore such decisions have to be made by a human. Furthermore, automatic analysis will not be reliable until we have very large corpora (min. 20 million words), and even if we have such large corpora, automatic analysis of the closed word classes will be difficult. In

[6] The figure in (9) is taken from Eide (2005:Chapter 2, figure 6). Here, I have only included the modals that occur in the Norwegian original.

the German *Wortschatz* project[7] (cf. Biemann and Quasthoff 2006) it has been shown that to obtain reliable results by automatically finding correspondences in parallel corpora at least 10.000 sentence pairs are needed (cf. Biemann 2005 and Cysouw, Biemann, and Ongyerth 2006).

Chris Biemann (personal communication) has informed me that a parallel corpus of 500 sentences will only give results where 35% of low frequency words (freq>0) are correctly analyzed and approx. 50% on freq>20 (note that only one modal, *kunne*, occurred more than 30 times in the Norwegian original, cf. (10) below). Biemann has tested four different sizes of corpora: 300 sentences (which did not give any results at all), 500, 3.000, 10.000, and 1 million. The figures for low frequency words become more reliable as the size of the corpora increases. As soon as the corpus contains 3.000 sentences, 40% of freq>0 are correctly analyzed, and up to approx. 70% of freq>20.

3. The results

Six Norwegian modals occur in the first two chapters of *Sofies verden*.

(10) Norwegian modals in *Sofies verden* (number of occurrences)
kunne (44), *skulle* (28), *måtte* (22), *ville* (21), *burde* (6), *få* (3)

In the following sections, results for each of the modals will be presented separately, in two different ways, in a table as in (11), presenting the possible correspondents, and as a figure as in (12). The table lists the combinations of correspondents according to their occurrence in the alignment blocks:

(11) Table X. Correspondents of *y*

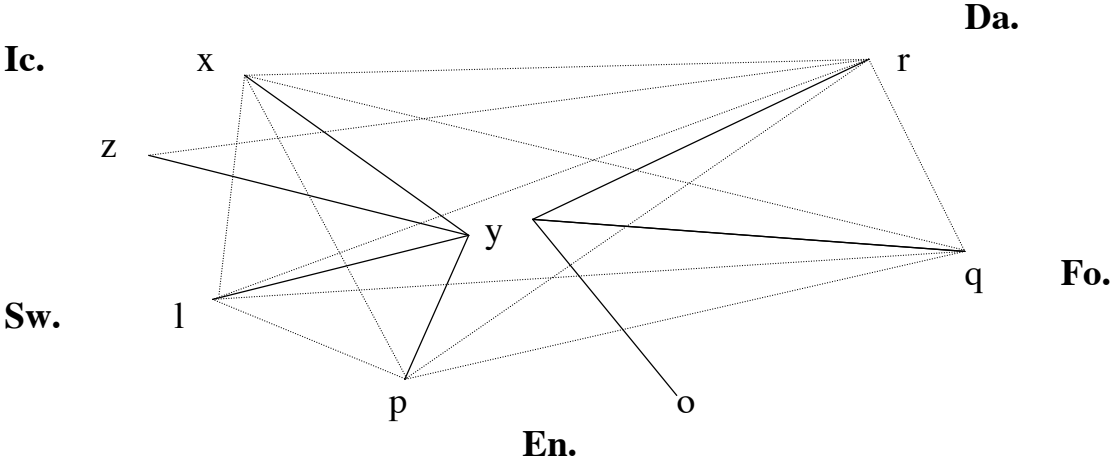
Nob.	Ic.	Fo.	En.	Da.	Sw.
y	*x*	*q*	*p*	*R*	*l*
	z			*R*	
			o		

The table in (11) shows that in some alignment blocks, *y* has the correspondents *x*, *q*, *p*, *r*, and *l*. In other alignment blocks, *y* has the correspondents *z* and *r*, and in yet other alignment blocks, *y* has the correspondent *o*. When these correspondences are fed into the database, a pattern that links each of the modals to one another, similar to the figure in (12) below, appears. The solid lines indicate the correspondences between the Norwegian original and the translations; the dashed lines indicate the

[7] Go to http://wortschatz.uni-leipzig.de/ for further information on the *Wortschatz* project.

resulting correspondences between the translations. The Norwegian original is always in the middle of the figure:

(12) Figure X. Links between correspondences of *y*



This is how to read the figure: From the solid lines, we see that *y* from the original corresponds to *p* and *o* in the English translation and to *z* and *x* in the Icelandic translations. As the dashed lines show, however, there is only correspondence between *p* in the English translation and *x* in the Icelandic one. There is no correspondence between *o*, *z*, and *x* and there is no correspondence between *p* and *z*.

Even though the method was limited to finding correspondences between Norwegian and the other languages, we get correspondences between e.g. Danish and Icelandic for free. As already mentioned, the data set is very small. Therefore, care should be taken *not* to regard the results that I report in the following sections to be conclusive. The results should only be taken as an indication of what things look like, an indication that might be used in further extraction of modals from parallel Scandinavian corpora.

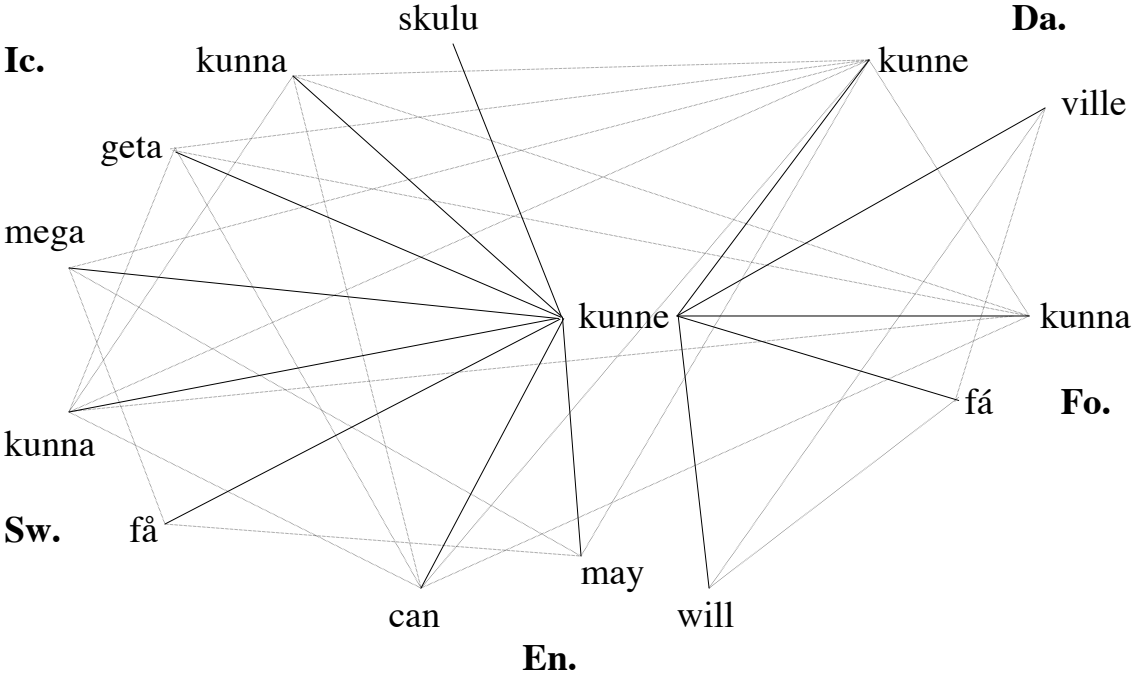
3.1 Kunne

Norwegian *kunne* is either translated as *geta*, *mega*, *skulu,* or *kunna* in Icelandic. In Faroese, either *kunna* or *fá* is used. Likewise, in Swedish, it is *kunne* and *få*. In the English translation, *can*, *will,* or *may* are used whereas Danish only uses *kunne*. Table 4 shows the possibilities, whereas Figure 1 in (14) shows how the correspondences of *kunne* are linked together:

(13) Table 4. Correspondents of *kunne*

Nob.	Ic.	Fo.	En.	Da.	Sw.
kunne	*geta*	*kunna*	*can*	*kunne*	*kunna*
	kunna	*kunna*	*can*	*kunne*	*kunna*
	mega		*may*	*kunne*	
	mega		*may*		*få*
	kunna				
	skulu				
	mega				
		fá	*will*	*ville*	
			may		
			will		
					få

(14) Figure 1. Links between correspondences of *kunne*



3.2 Skulle

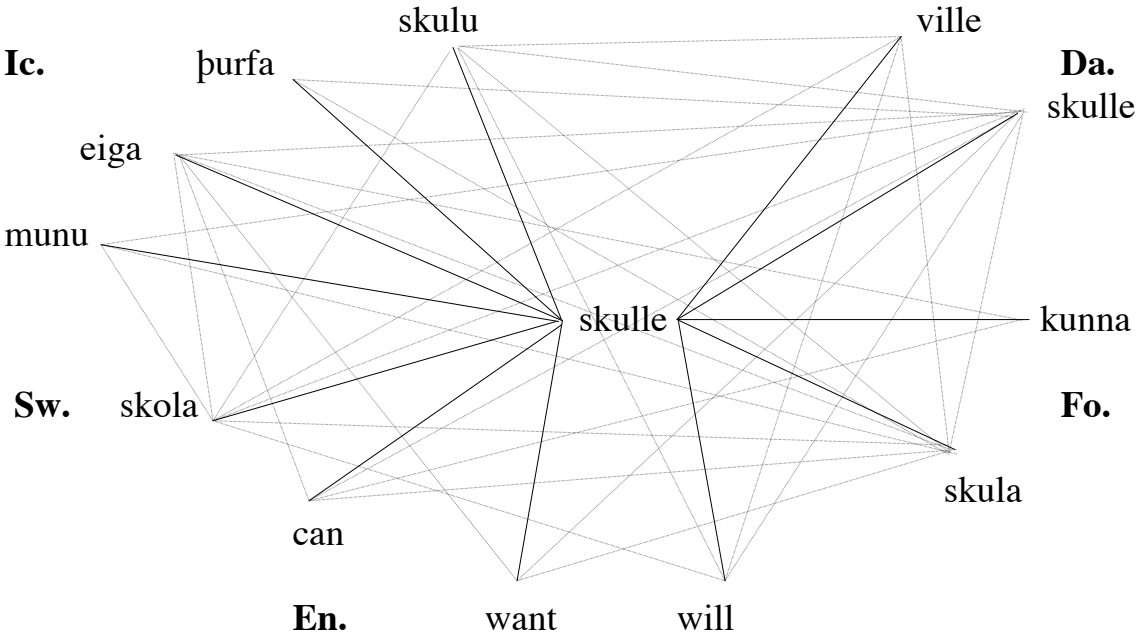
Norwegian *skulle* is translated into four different modals in the Icelandic translation of *Sofies verden*, *skulu*, *munu*, *eiga*, and *þurfa*. In the Faroese version, it is either translated with *skula* or *kunna*. English either uses *will*, *can*, or *want*; Danish uses either *skulle* or *ville*. In Swedish (which only has half of the sentences the other languages have) only *skola* occurs.[8]

(15) Table 5. Correspondents of *skulle*

Nob.	Ic.	Fo.	En.	Da.	Sw.
skulle	*skulu*		*will*	*ville*	*skola*
	munu	*skula*		*skulle*	*skola*
	munu				
	eiga	*skula*	*can*	*skulle*	
	eiga	*skula*		*skulle*	*skola*
	eiga	*skula*		*skulle*	
	eiga			*skulle*	*skola*
	eiga	*skula*	*want*	*skulle*	
	eiga	*kunna*	*can*	*skulle*	
	skulu	*skula*		*skulle*	
	skulu	*skula*		*ville*	
	þurfa	*skula*		*skulle*	
		skula		*skulle*	*skola*
		skula		*skulle*	
		skula			*skola*
			will	*skulle*	
				ville	*skola*
				skulle	

[8] It has been pointed out to me that Swedish *skola* and *böra* are very archaic and not used anymore as the infinitive of the modal verbs, which have the form *skulle* and *borde* in the past tense. However, *skola* and *böra* are both listed as the infinitive in the Swedish reference grammar that I used (Holmes and Hinchliffe 2003:257).

(16) Figure 2. Links between correspondences of *skulle*



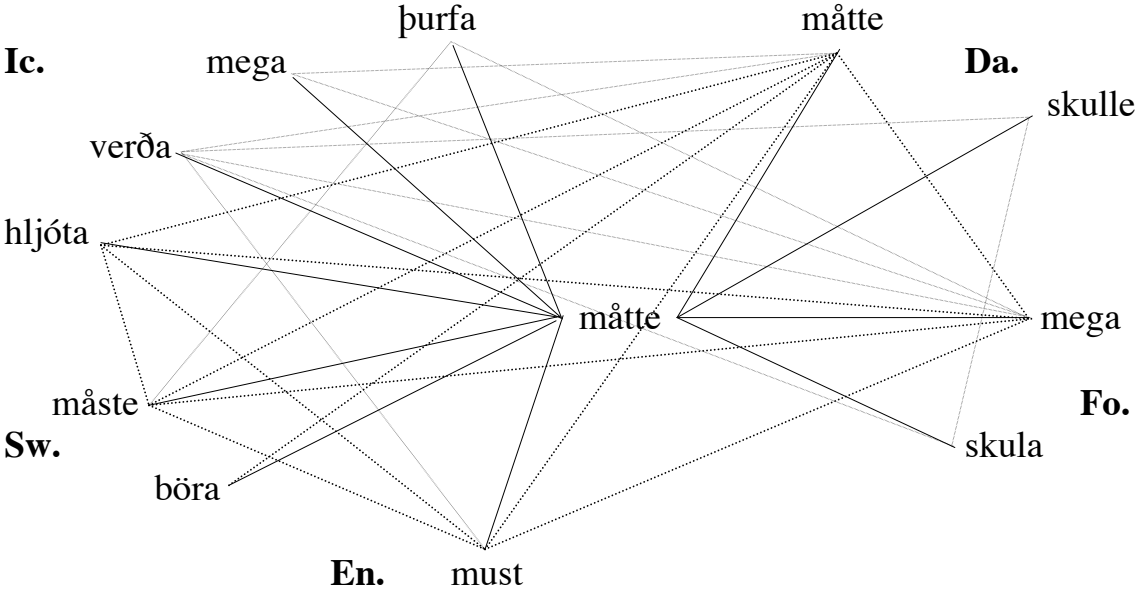
3.3 Måtte

Like *kunne* and *skulle*, Norwegian *måtte* is translated into four different modals in the Icelandic translation, *hljóta*, *verða*, *mega*, and *þurfa*. Faroese either uses *mega* or *skula*, while English uses *must*. In the Danish translation either *måtte* or *skulle* is used and in the Swedish translation it is either *måste* or *böra*.

(17) Table 6. Correspondents of *måtte*

Nob.	Ic.	Fo.	En.	Da.	Sw.
måtte	*hljóta*	*mega*	*must*	*måtte*	*måste*
	hljóta	*mega*		*måtte*	*måste*
	hljóta	*mega*		*måtte*	
	hljóta			*måtte*	
	verða	*mega*	*must*	*måtte*	
	verða	*mega*		*måtte*	
	verða	*mega*			
	verða	*skula*		*skulle*	
	mega	*mega*		*måtte*	
	mega			*måtte*	
	þurfa	*mega*			*måste*
		skula		*skulle*	
				måtte	*böra*

(18) Figure 3. Links between correspondences of *måtte*



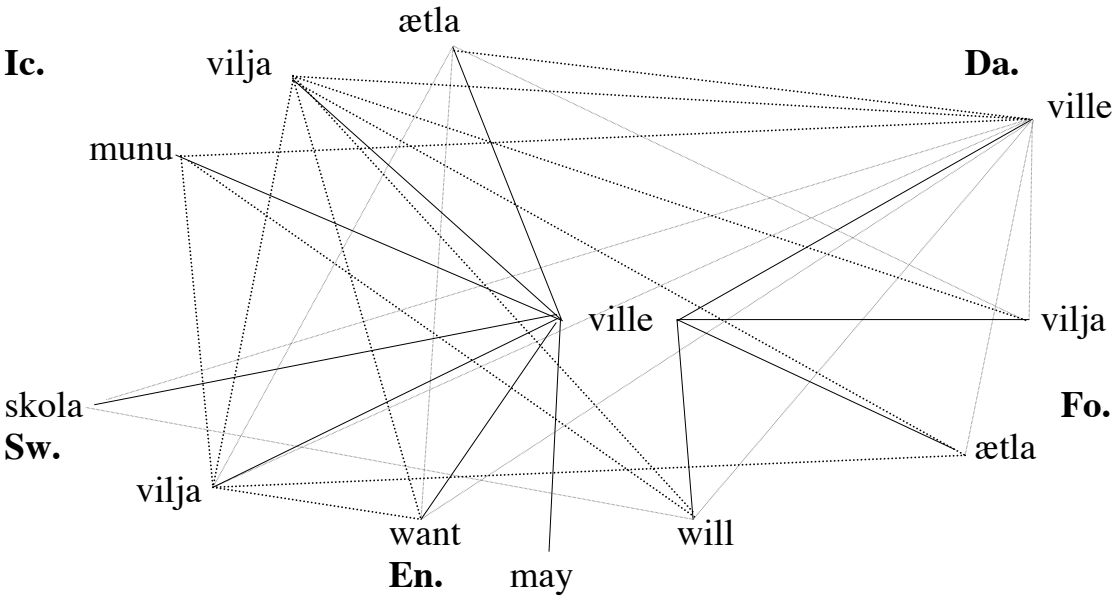
3.4 Ville

Norwegian *ville* is translated into three different modals in the Icelandic translation, *munu*, *vilja*, and *ætla*. Faroese either uses *vilja* or *ætla*, while English uses *will*, *may*, or *want*. *Ville* is translated into *ville* in the Danish translation and into *skola* or *vilja* in the Swedish translation.

(19) Table 7. Correspondents of *ville*

Nob.	Ic.	Fo.	En.	Da.	Sw.
ville	*munu*		*will*	*ville*	
	munu		*will*		
	munu			*ville*	
			will	*ville*	*skola*
			will	*ville*	
				ville	
			may		
	vilja	*vilja*	*will*	*ville*	
	vilja	*vilja*	*want*	*ville*	
	vilja	*ætla*		*ville*	*vilja*
	ætla	*vilja*	*want*	*ville*	*vilja*
	munu		*will*		*vilja*
	munu		*will*	*ville*	

(20) Figure 4. Links between correspondences of *ville*



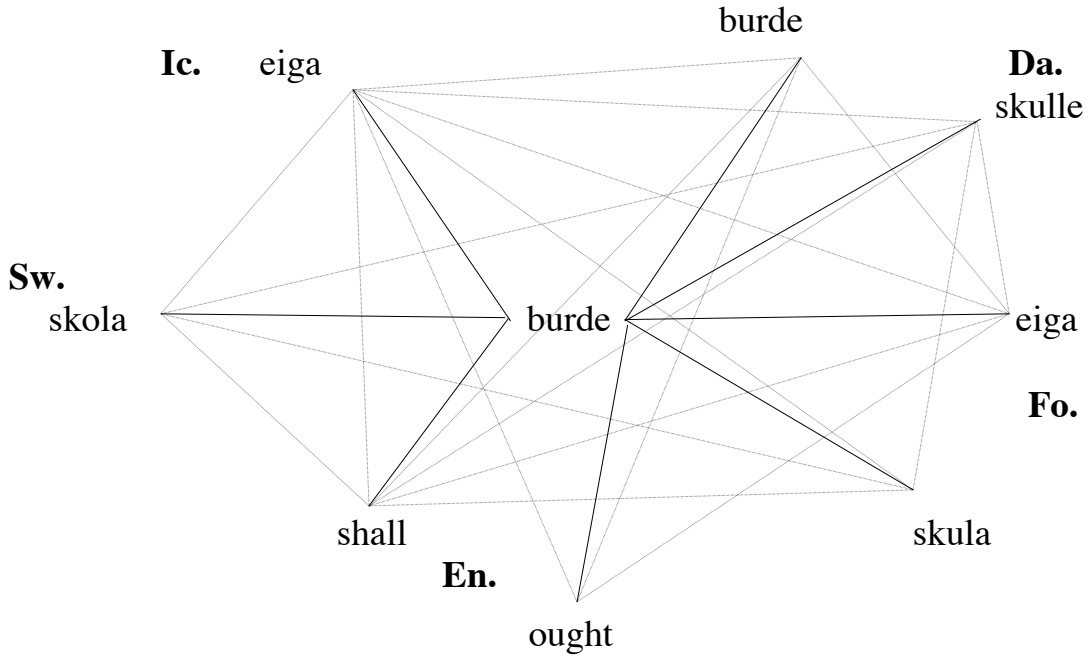
3.5 Burde

The six occurrences of Norwegian *burde* in the original are translated into *eiga* in the Icelandic version, *eiga* or *skula* (4/2) in the Faroese version, *ought* or *shall* (2/3) in English, *burde* or *skulle* (5/1) in Danish, and *skola* (1) in the Swedish translation.

(21) Table 8. Correspondents of *burde*

Nob.	Ic.	Fo.	En.	Da.	Sw.
burde	*eiga*	*eiga*	*ought*	*burde*	
	eiga	*eiga*		*burde*	
	eiga	*eiga*	*shall*	*burde*	
	eiga	*skula*	*shall*	*burde*	
	eiga	*skula*	*shall*	*skulle*	*skola*

(22) Figure 5. Links between correspondences of *burde*



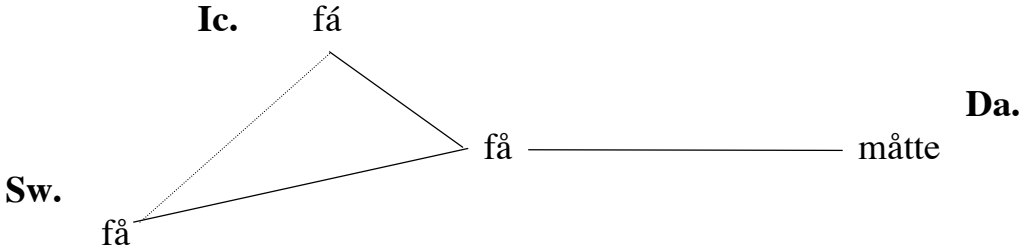
3.6 Få

The modal *få* occurs three times in the Norwegian original. Two of those have correspondences in Danish (*måtte*), Icelandic (*fá*), and Swedish (*få*).

(23) Table 9. Correspondents of *få*

Nob.	Ic.	Fo.	En.	Da.	Sw.
få	*fá*				*få*
				måtte	

(24) Figure 6. Links between correspondences of *få*



Although there is no correspondence between Danish *måtte*, Icelandic *fá*, and Swedish *få* in (24), it might as well be possible that there is a correspondence between the three verbs. We just cannot see it here because of the small data set.

4. Conclusions

The experiment shows that it is possible to generate a lexicon of modal verbs from a parallel corpus. It is possible to use automatic methods to extract and align the data. Here the alignment has been done on the

sentence level, but it is possible to align on the word level as well. At the moment it is not possible to automatically disambiguate epistemic modals from deontic modals.

It turned out not to be necessary to find the correspondences between the translations by looking at each language individually. On the contrary, it was sufficient to find the correspondents of the Norwegian original. When the corresponding sentences had been aligned, the correspondents between the translations became clear as we saw in the tables.

Since the data set is small, the results from the experiment should not be considered to be conclusive. Elsewhere, it has been shown that for reliable results, corpora with at least 3.000 sentences are needed. Such corpora do not exist for all of the Scandinavian languages.

References

Biemann, Christian. 2005. 'The Wortschatz Project: Language independent methods for enriching corpora,' talk given at the University of Oslo, October 11, 2005.

Biemann, Cris and Uwe Quasthoff. 2006. 'Dictionary acquisition using parallel text and co-ocurrence statistics,' in S. Werner (ed.) *Proceedings of the 15th NODALIDA*, Joensuu, 22-29.

Cysouw, Michael, Christian Biemann and Matthias Ongyerth. 2006. 'Using Strong's Numbers in the Bible to test and Automatic alignment of Parallel texts,' in Michael Cysouw and Bernhard Wälchli (eds.) *Parallel Texts: Using translational equivalents in linguistic typology*. Special issue of *Sprachtypologie and Universalien-Forschung*, 66-79.

Eide, Kristin M. 2005. *Norwegian Modals*. Studies in Generative Grammar 70, Mouton de Gruyter, Berlin.

Faarlund, Jan Terje, Svein Lie and Kjell Ivar Vannebo. 1997. *Norsk referansegrammatikk*. Universitetsforlaget, Oslo.

Holmes, Philip and Ian Hinchliffe. 2003. *Swedish: A Comprehensive Grammar*. Second edition. Routledge, London.

Tiedemann, Jörg. 2003. *Recycling Translations - Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing*. Doctoral Thesis, Studia Linguistica Upsaliensia.