


NORDISKE STUDIER I LEKSIKOGRAFI

Titel:	A corpus based method for a diachronic study of the central vocabulary of New Norwegian	
Forfatter:	Daniel Ridings & Oddrun Grønvik	
Kilde:	Nordiska Studier i Lexikografi 11, 2012, s. 524-533 Rapport från Konferens om lexicografi i Norden, Lund 24.-27. maj 2011	
URL:	http://ojs.statsbiblioteket.dk/index.php/nsil/issue/archive	

© Nordisk forening for lexicografi

Betingelser for brug af denne artikel

Denne artikel er omfattet af ophavsretsloven, og der må citeres fra den. Følgende betingelser skal dog være opfyldt:

- Citatet skal være i overensstemmelse med „god skik“
- Der må kun citeres „i det omfang, som betinges af formålet“
- Ophavsmanden til teksten skal krediteres, og kilden skal angives, jf. ovenstående bibliografiske oplysninger.

Søgbarhed

Artiklerne i de ældre Nordiske studier i lexicografi (1-5) er skannet og OCR-behandlet. OCR står for 'optical character recognition' og kan ved tegngenkendelse konvertere et billede til tekst. Dermed kan man søge i teksten. Imidlertid kan der opstå fejl i tegngenkendelsen, og når man søger på fx navne, skal man være forberedt på at søgningen ikke er 100 % pålidelig.

A corpus based method for a diachronic study of the central vocabulary of New Norwegian

Daniel Ridings & Oddrun Grønvik

This article describes how a monitor corpus can be created from existing corpora where the texts are annotated with basic bibliographical information.

The focus is on the core vocabulary and how to document diachronic change. The implementation uses simple, readily available techniques from corpus linguistics.

A norm is defined by isolating a small vocabulary unlikely to change from text to text, consisting of the most frequent words. These words are the building blocks of language – function words and some highly frequent verbs and nouns, words essential to producing grammatical sentences. The relative frequencies of these words from one text collection to another will show minor deviations. This slight deviation is used to specify the norm. It is the ‘wobble room’.

Sub-corpora are then created of the corpus sections to be compared. Relative frequencies are produced for every word in both subsets. When the relative frequency of a word from one subset deviates more than the above-mentioned wobble room compared to the same word in the other chronological subset, something has happened. A word may be moving in or out of the language or texts in one collection may deviate drastically from texts in the other, or an event may have raised an otherwise obscure technical term to the front page.

Key words: monitor corpus, relative frequencies, central vocabulary, New Norwegian, corpus linguistics

Corpora are now main-stream for any dictionary project with its own editors. The methodology was set out in the COBUILD dictionary project (Sinclair 1987) and has spread throughout the world. If a dictionary project does not have a corpus, it wants one.

Another trend that goes back to John Sinclair and COBUILD is the characterization of a corpus (Sinclair 1996). Corpora have their typologies. They can be domain specific, useful in the creation of terminologies, balanced, reference,

attempting to provide a gold standard for a language, and monitor, documenting the changes in language over time. In order for a corpus to be really useful for an ambitious dictionary project there is one characteristic that subsumes most of the others. The corpus should be big. If the corpus is large in number of tokens or documents, but skewed by containing for instance too many newspaper texts, it is not big enough. It should preferably be augmented by adding other text types, rather than by cutting newspaper texts and thereby diminishing the size. If there are electronic texts easily available and they are not in the corpus, then the corpus probably needs to be bigger. Scandinavian languages are not spoiled by a wealth of digital texts like some other languages, such as English.

In the mid-nineties, corpora tended to land around 30-40 million words (Sinclair 1996). It was a comfortable size. Online concordances could be generated dynamically and the resulting concordance lines handled fairly comfortably. It was certainly easier than working one's way through the texts and creating archives of dictionary slips.

Computers continue to grow in power, storage continues to drop in price, more and more text is available electronically. Corpora grow and grow. They become so large that they cannot be handled comfortably by traditional means. A lexicographer can only digest so many concordance lines. The known facts about language get reinforced by more and more evidence, but the interesting subtle changes and aspects get drowned out in the wealth of material.

Norsk Ordbok 2014

The beginnings of the New Norwegian corpus go back to the early 90-ies. Two factors played an important role in the beginning. The first factor was an ambitious project at the University of Oslo 1991-1997, the *Documentation Project*. Its goal was to provide an extensive information system for the Norwegian language and culture. It eventually involved several other universities in Norway, but in the present context, it was the project's focus on providing electronic research material for language studies in general and lexicography in particular that makes it important (Ore et al 1998: 89 ff.).

The second factor was a chance opportunity to test out the importance and usefulness of corpora for poorly documented languages with a short history of written literature. In 1992 representatives from the University of Norway, Christian-Emil Smith Ore and Oddrun Grønvik and two representatives from Göteborg University, Martin Gellerstam and Daniel Ridings, combined their efforts in a NUFU and SIDA financed project at the University of Zimbabwe in

Harare. The project was called the *African Languages and Lexicon Project* (ALLEX). Corpora already had a solid position in Swedish lexicography. The methodology spread to ALLEX and, indirectly, to Norsk Ordbok 2014 as from 2002. The Documentation Project had digitalized many texts for Norwegian and at least two invaluable resources for New Norwegian: Garborg's and Vinje's collected works along with some key smaller texts by Ivar Aasen.

In 2002 one of the authors of this paper, Daniel Ridings, joined the staff of the *Unit for Digital Documentation* at the University of Oslo with the explicit task of building up a New Norwegian corpus for the national dictionary project, Norsk Ordbok 2014 (Ridings 2005).

The norm

The basic idea of this presentation is that there is a core vocabulary that does not change much from text type to text type. The assumption that the core vocabulary is fairly stable, allows us to turn to the vocabulary that actually does change. These changes can be due to a text type or terminological domain, or the change can be due to time, that is, we can investigate the vocabulary diachronically. This is an attempt to describe how Sinclair's monitor corpus can be implemented.

It goes without saying that any vocabulary that does not change significantly from a scientific text to a non-fiction text is not going to consist of content words. Content words will reflect the domain and text type, and the fact that they do change significantly from one domain to another is used in various techniques, from extracting terminology to the automatic production of keywords and summaries. Function words, the small building blocks of language, are the ones that will be found in any text or transcription of speech.

Another pointer to where we can find the words at the very core of a language comes from experience with part of speech tagging. A common tactic in part of speech tagging is to look up each word of a text in a digital lexicon, a lexicon that consists of a word forms and a list of analyses that a given form can have. That is the easy part, if there is a lexicon available. The hard part, the part where most energy is spent, is disambiguating the tags when a word form can have two analyses or more. *Record* can be a noun and *record* can be a verb. The main task of a part of speech tagger is to decide between the alternatives and assign the correct analysis to an ambiguous word based on the surrounding context.

A few words about the digital lexicon – the reference point for part of speech tagging. The more words there are in the lexicon, the better, up to a point. The

Norsk Ordbok 2014 New Norwegian corpus is fairly large, by any standards. At present, the autumn of 2011, it consists of 87,767,084 running tokens. That is the size of the corpus. Of these almost 88 million tokens, 1,863,524 are unique combinations of a word and a part of speech description. The remaining tokens need disambiguation. The token *kasta*, for example, can be a past tense form, a perfect participle, an infinitive or the definite form of *kaste* in the singular, to name a few. In order to have a lexicon to cover all the analyses of the words in the corpus, the lexicon would have to be very large. In practice, lexica for part of speech tagging are kept much smaller by making certain educated guesses. One such would be that all words with a final *-ane* can be assumed to be masculine plural nouns in the definite form until shown to be otherwise, by the context.

Those who have worked with part of speech tagging know that a lexicon does not need to be unreasonably large. They also know that a good lexicon will do most of the work for them. The lexicon available to the Norsk Ordbok 2014 New Norwegian corpus is the New Norwegian section of *Norsk ordbank* (the Norwegian Word Bank), which has close to 100 000 lemma with paradigms.

A lot of effort is put into raising the precision of part of speech tagging from 94% to 96%, but a good lexicon will reach somewhere between 85 – 90% all by itself, without fancy algorithms. Why is this?

As large as the NO 2014 New Norwegian corpus is, a lexicon would only have to contain 53 words in order to tag 40% of the whole corpus. In other words, 53 words are all that is needed in the lexicon in order to tag over 35 million words of the corpus. 172 words in the lexicon will cover 50% of the corpus, a little over 43 million words, and 592 words in the lexicon will cover 60% of the corpus. It feels like a safe hypothesis to assume that the 53 most frequent words do belong to the core vocabulary. If one million running words are added to the corpus, the assumption is that these 53 words will not show much deviation with regard to relative frequencies. The following is a list of what these 53 words were in 2009. The relative frequencies are per 100,000 words. For example, one can expect to find the word *og* 2,940 times per 100,000 words of the corpus.

og	2940.0	med	994.3	har	631.9
i	2401.7	ein	915.1	eg	618.3
det	2204.6	for	875.8	seg	558.7
er	1378.2	var	843.7	men	557.5
som	1362.9	at	837.4	om	471.5
til	1250.3	av	826.0	ho	462.4
han	1226.9	ikkje	811.3	eit	444.0
på	1131.9	å	791.7	hadde	367.1
dei	1001.4	den	662.7	ei	358.8

frå	323.1	skal	237.9	alle	165.5
vi	302.4	ut	229.1	andre	163.7
så	302.3	paa	222.3	meg	162.8
du	277.1	vart	218.8	her	159.1
der	271.2	vil	194.7	noko	152.7
kan	263.3	dette	193.1	mot	152.6
so	251.3	eller	174.0	aa	149.2
no	248.6	berre	173.2	over	147.5
etter	240.5	kom	170.6		

In May 2009 an opportunity arose to test the assumptions made here. The corpus was already quite large with a wide selection of texts from across the spectrum of New Norwegian literature. A whole year volume of *Dag og Tid* from 2006 was acquired and the question arose: what happens when you add one million words from a specific genre?

Before the new texts were added to the corpus, a list was made all the words in the corpus with their relative frequency. This list was then recreated after the one million words of newspaper texts were added. Then the two lists were compared. There is bound to be some deviation between the relative frequencies for a given word in the two lists. The problem was to figure out how much deviation was significant.

It was known that there were some words in the top of the frequency list that would be affected radically. These were words that reflected the spelling norm from the 1800's and from before 1938, when a major spelling reform was introduced. Since the texts being added were all from after the spelling reform, it was reasonable to assume that the older words would display a *worst case scenario* when it came to a deviation of their relative frequencies before and after. *Paa* and *aa* are examples of such words in the list above. The following list is slightly larger than the one above. It contains the top 76 words. This time the numbers do not represent the relative frequencies, but a measure, in per cent, of how much the relative frequencies differ from what they were before the one million words of modern newspaper texts was added. Some, in particular words from earlier orthographical norms, have, as expected, decreased (-) in relative frequency by a large measure. Others have differed by much smaller measures. The words are sorted in descending order according to how many percentage points they differ from the relative frequencies they had before the new addition to the corpus was made.

aa	-	3.514	um	-	3.385	henne	-	2.528
paa	-	3.512	mange	+	2.567	har	+	2.469
so	-	3.421	år	+	2.552	om	+	2.322

sa	-	1.941	på	+	1.018	meg	-	0.427
seier	+	1.917	dette	+	1.013	alle	+	0.420
å	+	1.853	vore	+	0.995	over	+	0.403
frå	+	1.638	i	+	0.945	to	+	0.366
gjekk	-	1.628	for	+	0.937	så	+	0.355
av	+	1.564	ei	+	0.926	slik	+	0.355
ho	-	1.529	som	+	0.867	inn	-	0.310
er	+	1.517	at	+	0.866	opp	+	0.274
hadde	-	1.465	sjølv	+	0.846	med	+	0.251
vera	-	1.444	enn	+	0.828	ikkje	+	0.238
då	+	1.333	den	+	0.790	vart	+	0.216
kjem	+	1.285	seg	-	0.775	skal	-	0.215
når	+	1.281	til	+	0.746	me	+	0.207
han	-	1.267	eg	+	0.746	og	-	0.204
må	+	1.259	eller	+	0.732	etter	-	0.194
du	-	1.254	men	+	0.663	det	+	0.151
kan	+	1.251	alt	-	0.638	fram	-	0.132
også	+	1.236	meir	+	0.626	ved	-	0.085
ein	+	1.220	noko	+	0.585	berre	+	0.041
hans	-	1.202	ha	+	0.573	kunne	+	0.036
kom	-	1.193	vi	+	0.566	mot	+	0.020
eit	+	1.152	andre	+	0.554	dei	+	0.019
denne	+	1.148	fekk	-	0.553	vil	-	0.018
få	+	1.044	der	-	0.551	ut	+	0.009
var	-	1.034	her	-	0.549	blir	+	0.001
kva	+	1.032	no	-	0.536			

As expected, it can be seen that *aa*, *paa*, *so*, and *um* have fallen the most in relative frequency. They all belong to an older orthography and will only be found in modern texts containing citations from older texts. Their frequencies have decreased from 3.3% to 3.5%. On the other end of the spectrum, the spectrum being the 76 most frequent words, there are words such as *blir*, *ut*, *vil* that have hardly changed at all. This is taken as evidence that there is a stable kernel of words in New Norwegian that are not dependent on text type or domain. One million words of newspaper text could be expected to produce a tangible slant on the vocabulary of corpus, but that is not what is found in these words. On the average, the most frequent words displayed a variation of relative frequency of 1.1%.

Synchronic study

This figure was then used for a synchronic study, to demonstrate how a corpus can be used to identify good candidates for a bilingual or learner's dictionary of New Norwegian. As mentioned above, there are over 87 million words, or more accurately *tokens*, in the corpus. Many tokens are productive compounds, numbers, names and the like. Such words will not automatically become candidates for a dictionary, no matter how frequent they are. A simple frequency lists in descending order will still leave a substantial amount of work for the lexicographers to filter the list down to a manageable vocabulary. What would the principles be? Would they be principles that could be taught to others and principles that could be objectively applied or would they be based on subjective intuition?

It was then demonstrated how one can, by using the figures above, and settling on a maximum variation of 2.5%, isolate a group of words in the whole corpus that display the same amount of stability as the 75 most frequent words.

This criterion filters out the low frequency words. If there was a hapax in the corpus before the newspaper texts were added, and that word was found one more time in the newspapers, its increase in relative frequency would be 100%.

Two lists of words from the corpus were produced. The two lists represented the corpus vocabulary together with relative frequencies from before and after the addition of new texts. A collection of approximately 25,000 word forms (25,597) was isolated. These word forms display the same measure of stability as was defined by the top most frequent words, the absolute core vocabulary. This list would provide a very good starting point for the vocabulary of a basic dictionary of New Norwegian, based on objective criteria.

Diachronic extension

This method can be used in many circumstances where one can think in terms of a sub-corpus. The corpus prior to the addition of new texts is a sub-corpus of the whole corpus after the texts have been added.

Each text of the New Norwegian corpus has been assigned a number of attributes. A title is one such attribute. An author is another and the date of publication is yet another. Every single text has the year of publication associated with it. This last attribute, the date, provides what is necessary to perform diachronic investigations of the vocabulary.

It is possible to create sub-corpora based on publication date. Frequency data for all texts before a given year, 1970, for example, can be created as a sub-corpus. Then frequency data for the same texts with the addition of those from

1970–1979 can be created and compared to the frequency data for the pre-1970 texts analogous to the comparison that was made above when one million words of newspaper texts were added to an existing corpus. This time, however, the comparison is not made between existing and new texts, but between subsets of the same corpus. This enables a diachronic investigation of the vocabulary without having to add the texts in chronological order.

The New Norwegian corpus is stored in an Oracle database in order to be integrated into the editing and formatting software that has been produced for the lexicographers by the *Unit for Digital Documentation* at the University of Oslo. There are four main tables:

1. One table, GRAPHWORD, for the word-forms found in the corpus.
2. One table, OCCURENCES, that ties a word-form to a position in a text, one record for each position. If a form occurs 25 times, there will be 25 rows for that form in this table.
3. A table for the text, TEI-markup and all.
4. A table for textual features, DOCUMENTS, containing a unique text number, the abbreviation used for a text in concordances, the year of publication and some other features of no interest here.

It is not necessary to create copies of the corpus when one wants to work with various sub-corpora. The following SQL query, written by Christian-Emil Smith Ore, will create a frequency list of all texts before the year 1970:

```
SELECT
    t4.word, t4.id, t3.datum, count(*) z
FROM
    occurrence t1, text t2, document t3, wordtype t4
WHERE
    (t1.t_id = t2.id) and
    t1.w_id = t4.id and
    (t2.document_id = t3.id) and
    (t3.datum < '1970')
GROUP BY
    t4.word,
    t4.id,
    t3.datum
ORDER BY z DESC
```

It takes a few seconds and a new corpus is not created, only frequency information for all texts before 1970. This frequency list is set aside and the relative frequencies are calculated.

A similar query, changing the date in the query to include texts published earlier than 1980, will create a new frequency list. This list will include all the words in the first list in addition to data for all the words from the next decade as well. This is the equivalent of having a corpus of texts published before 1970 and then adding new texts from the following decade. This list is also set aside and the relative frequencies are calculated.

These two data-sets, word lists with relative frequencies, can now be processed using the techniques described for synchronic processing above. However, instead of looking for *stable* words, a diachronic study might be more interested in the words that have *entered* written language, that is, words that display a greater increase in relative frequency than stable words are expected to have. Or one could look for words that appear to be *leaving* the language, that is, words that display a greater negative change in relative frequency than stable words are expected to have. Such a process can be described as monitoring the language. The corpus is the same, but the techniques used to process it create a monitor corpus.

In principle, one could do this processing of the corpus from the earliest texts to the most recent texts in ten year increments and study the changes in vocabulary, spelling norms and inflectional patterns across time. It would be interesting to see how rapidly the various orthographical recommendations took root.

The techniques described here could also be extended for the extraction of objectively based keywords for a text. It could be argued that a text that strongly reflects a certain subject or domain, will inevitably display differences in the vocabulary. Syntactic structures remain relatively stable across domains, but not the vocabulary. One could add a strongly domain biased text to a general language corpus and extract the words that have increased in relative frequency. In such a text, it is not unreasonable to assume that such words would point to what the text is *about*. This technique allows the text itself to declare what it is about. It is notoriously difficult for a number of individuals to assign the same keywords to the same text. The keywords individual experts chose are based on subjective criteria and the subjective opinions of several individuals do not necessarily coincide. The value of the resulting keywords is accordingly.

One could also compare two authors to each other, Garborg and Vinje, for example. Or one could take a text with an unidentified author and compare it to one by a known author. If the identified author has some idiosyncrasies when compared to others and if the unidentified work displays some of the same behavior, then the techniques could be used to point the way towards a closer inspection.

The possibilities are many. We hope that this paper has described the process in enough detail for it to be reapplied in installations with a different infrastructure.

LITERATURE

- Ore, C.-E. S. og Kristiansen, N. 1998: Sluttrapport 1992-1997. Dokumentasjonsprosjektet. Universitetet i Oslo.
- Ridings, Daniel, 2005: Nynorskorpuset vid Norsk Ordbok 2014: Integreering med redaktions-arbete. In: Nordiske studiar i leksikografi 8, pp. 315-325.
- Sinclair, John (ed.), 1987: Looking up: An account of the COBUILD Project in lexical computing. London and Glasgow.
- Sinclair, John, 1996: EAGLES, Preliminary recommendations on Corpus Typology, <http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html>.

The ALLEX Project 1991 – 2006: See website at <http://www.edd.uio.no/allex/>
 Norsk Ordbank 1997-: See <http://www.hf.uio.no/iln/om/organisasjon/edd/forskning/norsk-ordbank/>

Daniel Ridings
 daniel@dlridings.se

Oddrun Grønvik
 NO2014, Universitetet i Oslo.
 oddrun.gronvik@iln.uio.no