

Constructing a Norwegian Academic Wordlist

Kristin Hagen, Janne Bondi Johannessen and Arash Saidi

Department of Linguistics and Scandinavian Studies, University of Oslo

Pb. 1102 Blindern, 0317 Oslo, Norway

E-mail: kristiha@iln.uio.no, jannebj@iln.uio.no, garash84@gmail.com

Abstract

We present the development of a Norwegian Academic Wordlist (AKA list) for the Norwegian Bokmål variety. To identify specific academic vocabulary we developed a 100-million-word academic corpus based on the University of Oslo archive of digital publications. Other corpora were used for testing and developing general word lists. We tried two different methods, those of Carlund et al. (2012) and Gardner & Davies (2013), and compared them. The resulting list is presented on a web site, where the words can be inspected in different ways, and freely downloaded.

Keywords: Norwegian, academic wordlist, method, academic corpus

1. Introduction

This paper presents the development of a Norwegian Academic Wordlist (AKA list) for the Norwegian Bokmål variety (for historical reasons, Norwegian has two written varieties, the other is called Nynorsk). Academic vocabulary is vocabulary that is more frequent in academic texts than in general texts. The individual words should be well represented in many kinds of academic genres, to avoid including subject-specific terminology. A brief description of academic vocabulary and why it is needed can be found in Section 2. To identify specific academic vocabulary we developed a 100-million-word academic corpus based on the University of Oslo archive of digital publications, described in Section 3. Other corpora were used for testing and for developing general word lists. We tried two different methods, those of Carlund et al. (2012) and Gardner & Davies (2013), Section 4. In Section 5, the lists are compared measuring coverage in two different test corpora. The resulting list is presented on a web site, where the words can be inspected in different ways, and freely downloaded (Section 6). The paper is concluded in Section 7.

2. Academic Vocabulary and the Need for a Norwegian AKA List

According to Gardner & Davies (2013:8), academic core words are “those that appear in the vast majority of the various academic disciplines”, in contrast to general high-frequency words “that appear with roughly equal and high frequency across all major registers of the larger corpus, including the academic register” and in contrast to academic, technical words “that appear in a narrow range of academic disciplines”.

Second language students and native students not used to academic language in their home and local environment face a challenge when meeting academia and its language. Unfortunately for such students in Norway, the nature of the Norwegian academic language has not been properly explored.

The situation is different for English: The first academic vocabulary lists were published in the 1970s (see references in Gardner & Davies 2013), the first representative, academic word list was published thirty years later (Coxhead 2000), and the latest approach was presented by Gardner & Davies in 2013.

We wanted to create a list of those academic words that people need to know independently of genre if they want to take full advantage of teaching and textbooks above school level, and to write academic texts. The list should be available for students who have Norwegian as a second language as well as students who have a non-academic background. It should also be offered to teachers and authors of text books for higher education.

3. The Academic DUO Corpus

Prior to our endeavours no general academic corpus existed for Norwegian, with the exception of a small bilingual corpus developed by the University of Bergen, the KIAP corpus, with 450 texts from three academic fields, on the one hand, and some subparts of large general language corpora (such as the 100-million-words Leksikografisk Bokmålskorpus (LBK)), on the other. We therefore had to develop an academic corpus of our own, which was greatly facilitated by the University of Oslo Library, which gave us permission to use their archive of digital publications (DUO) of master’s theses, doctoral dissertations, and journal publications.

After downloading the documents (in pdf format), converting them to text format, and cleaning out those that were not written in Norwegian Bokmål, we lemmatized and POS tagged the corpus, using the Oslo Bergen Tagger (Lynum et al. 2011, Johannessen et al. 2012). There are 3480 documents and approximately 100 million tokens in this version of the resulting academic DUO Corpus. The documents are from eight faculties, see Table 1. The corpus is larger than those used for Swedish and English, see Carlund et al. (2012), Coxhead (2000) and Gardner & Davies (2013).

Faculty	Departments within each faculty	Texts	Words
Faculty of Humanities	7	1236	44 706 060
Faculty of Educational Sciences	3	582	16 921 823
Faculty of Medicine	6	579	10 192 043
Faculty of Social Sciences (Economics, Sociology, Political Science)	6	542	13 908 454
Faculty of Mathematics and Natural Sciences	9	265	6 340 030
Faculty of Theology	4	114	3 545 930
Faculty of Law	4	111	4 660 631
Faculty of Dentistry	1	51	338 068
Sum	40	3480	100 613 039

Table 1: Numbers from the present DUO Academic Corpus

4. Constructing Norwegian Academic Wordlists (AKA-lists)

There are two main processes that must be used to identify academic words from an academic corpus (cf. Gardner & Davies 2013):

- exclude high-frequency words also found in non-academic language usage
- exclude terminology that has a high frequency in only some parts of the academic corpus

We describe two methods in this paper: The Gothenburg method (Carlund et al. 2012) for Swedish and the Gardner & Davies method (2013) for English.

4.1 The Gothenburg Method

Since Norwegian and Swedish are mutually understandable languages, we chose to try the same approach as Carlund et al. (2012). The Swedish Gothenburg method is quite similar to that of Coxhead (2000), but unlike Coxhead, it generates lemmas (lexical entries) and not word families (lemmas and their derivations). The academic word list was derived from the tagged and lemmatized DUO Corpus in four steps:

Keywords

A keyword is defined as a word that occurs with unusual frequency in a given text. Like Carlund et al. (2012) we ranked each word in the text according to its keywordness (following an algorithm described in Scott 1997) and set the first selection criterion to be a score of above 1.1. We used the web corpus NoWaC (Norwegian Web as Corpus), a 700-million-word corpus of Bokmål Norwegian, as a reference corpus. This method is meant to remove subject-specific terminology, but in actual fact had little effect.

Reduced frequency (RF) (range)

For each word the corpus was divided into a set of intervals based on the frequency of the said word, after which the intervals, in which this word occurred, were counted. This measure gave an indication of the extent to which a given word was spread out across the corpus. If a word had a high frequency in the corpus, but a low RF, it could be concluded that this word belonged to the

specialised vocabulary of some academic field. The way this was done, was to remove words that did not have a RF of at least 15 per million tokens in each of the sub-corpora (i.e., faculties), for details, see Carlund et al. (2012:§3.4). (Reduced frequency is related to term frequency-inverse document frequency.)

Removal of everyday words

Finally, using a stop list consisting of the most frequent words from a general language corpus, frequent words were removed. Such words would have a high frequency across the corpus and a high reduced frequency, which made them escape the first two steps above. Notice that this kind of stop list contains more general words and from different parts of speech, not only function words, as the classical stop lists often do.

Manual removal of “trash”

Some proper names like *Norge* (“Norway”) and *Oslo* survived the selection criteria above and had to be removed manually. So did text formatting abbreviations like *ii* and *a*.

Our resulting AKA lists clearly show that the choice of stop list and the size of the list have a big impact on the resulting AKA list. Coxhead (2000) used 2000 words from The General Service List (West 1953) as a stop list. Carlund et al. (2012) report that they removed the 1000 most frequent lemmas of everyday Swedish, calculated from the 1.1-million token corpus LäSBarT, a corpus containing children’s books and other easily read texts.

Unlike Coxhead and Carlund et al., we had no list of easy-to-read language, so we experimented with stop lists developed from the NoWaC corpus. We tested three stop lists from NoWaC: of 1000 words, 1500 words and 2000 words. An academic wordlist of the 500 most frequent words resulting from the Gothenburg method with each of the three stop lists was created and named AKA-1000, AKA-1500 and AKA-2000, respectively, see Figure 1 below. Notice that the first of these, AKA-1000, contains more general words than the other two. As the number of stop words increases, these more general words are removed.

The results from our experiments will be presented in the next section.

AKA-1000	AKA-1500	AKA-2000
1 kapittel (chapter)	1 kapittel (chapter)	1 kapittel (chapter)
2 betydning (meaning)	2 informant (informant)	2 informant (informant)
3 informant (informant)	3 analyse (analysis)	3 analyse (analysis)
4 rolle (role)	4 begrep (concept)	4 studie (study)
5 forståelse (comprehension)	5 intervju (interview)	5 problemstilling (thesis question)
6 analyse (analysis)	6 perspektiv (perspective)	6 uttrykke (express)
7 skille (differ)	7 studie (study)	7 tolke (interpret/explain)
8 begrep (concept)	8 definere (define)	8 relasjon (relation)
9 uttrykk (expression)	9 holdning (attitude)	9 oppfatning (understanding)
10 teori (theory)	10 faktor (factor)	10 aspekt (aspect)
11 metode (method)	11 problemstilling (thesis question)	11 funn (discovery)
12 intervju (interview)	12 uttrykke (express)	12 kontekst (context)
13 perspektiv (perspective)	13 omtale (mention)	13 tilnærming (approach)
14 studie (study)	14 beskrivelse (description)	14 struktur (structure)
15 definere (define)	15 tolke (interpret/explain)	15 teoretisk (theoretical)
16 oppfatte (perceive)	16 relasjon (relation)	16 bevisst (deliberate)
17 representere (represent)	17 undersøke (examine)	17 motsetning (contrast)
18 handling (act)	18 evne (ability)	18 fenomen (phenomena)
19 prege (mark)	19 oppfatning ((understanding)	19 belyse (illuminate)
20 holdning (attitude)	20 element (element)	20 likhet (similarity)

Figure 1: The 20 most frequent words from each of the three AKA lists. The words in dark grey print (blue), 11 words, are more general, and are found in the AKA-1000 list, due to this list's very short stop list. The words in light grey (yellow), 16 words, are found in the two AKA lists with the shortest stop lists. The words in black print, 33 words, are found in all three lists.

4.2 The Gardner & Davies Method

The G&D method has four steps. The first measure excludes high frequency words in the academic corpus in relation to a reference corpus. The other three exclude subject-specific words.

Ratio

To eliminate general high frequency words, a word in the academic DUO corpus must have a higher frequency than in the general corpus (here NoWaC). We experimented with a wide range of numbers here, and found that the most interesting results were words that were 2.2 - 2.6 times as frequent in the academic corpus compared to the general corpus. The 2.2 - 2.6 Ratio is a heuristic measure that we ended up with through experimentation. Notice that the ratio used by Gardner & Davies is 1.5 (Gardner & Davies 2013: 11), which shows that these measures are language and culture specific. There is no commonly accepted value for this measure.

Range

A word must occur with at least 30 or 40% of the expected frequency in the academic domains (here: at least 6 of the 8 faculties). For instance, if a word occurs 100 times in a corpus of 100,000 words, its frequency in relation to the corpus is $100/100.000 = 0.001$. Applying this to a sub-corpus of 10,000 words, the expected frequency would be $0.001 * 10.000 = 10$. In order for this word to pass the Range in this sub-corpus, the frequency of the word in said sub-corpus must be at least 3. We

experimented with values ranging from 20 to 60%, and experimentally found the range between 30 and 40% to give the best results.

Dispersion

Words in the academic corpus must have a Dispersion of at least 0.60. The Dispersion measure (see Julliland and Chang-Rodriguez 1964) is an indication of how "evenly" a word is spread in a corpus. The measure ranges from 0.01 (the word only occurs in a small part of the corpus) to 1 (even dispersion throughout the whole corpus). The Range measure ensures that a word is above a certain frequency threshold in most of the academic faculties, whereas the Dispersion measure ensures that the word is spread evenly throughout the corpus.

Discipline Measure

A word cannot occur more than 3 – 3.2 times the expected frequency (per million words) in any of the 8 faculties. Like Range and Dispersion this measure is designed to exclude discipline-specific words.

For each step in the methods above, then, we experimented with different input values that resulted in different lists. The choice between them was decided by testing coverage (see Section 5). Also from the Gardner & Davies method lists we had to remove some "trash" as described for the Gothenburg method.

5. Testing the two Methods

According to Nation (2001) the vocabulary of English

academic texts consists of almost 80 % high-frequency words, up to 9 % academic words and 5 % is subject-specific and technical words. It is important to keep this in mind.

We checked the AKA lists developed with the two methods against two test corpora:

- LBK Fiction (a 36.5 mill word sub-corpus of the Leksikografisk Bokmålskorpus (LBK corpus), a balanced corpus of Norwegian Bokmål
- The KIAP Corpus from University of Bergen (73 000 words; papers from economics, medicine and linguistics)

For the Gothenburg method our experiments showed that the results depend upon the stop list – both regarding size and origin. For the G&D method the word selection ratio appeared to be the most crucial.

Table 2 illustrates a selection of coverage measurements for five word lists, all of which were 750 words. Following Gardner & Davies (2013:18), we use a Random list as a measure of comparison. It was generated from a lemmatised sub corpus of the LBK corpus, by simply picking out 750 arbitrary words from the corpus. Observe that the list obviously contains many very frequent words. GM-1000 and GM-2000 are two lists generated by the Gothenburg method using stop lists (lists of most frequent words) of 1000 and 2000 words, respectively, extracted from NoWaC. G&D-2.6_0.3_0.6_3.2 and G&D- 2.2_0.4_0.6_3.0 show the G&D method with different input values (ratio, range, dispersion, and discipline) with NoWaC as a reference corpus.

	Random list	GM-1000	GM-2000	G&D-2.6_0.3_0.6_3.2	G&D- 2.2_0.4_0.6_3.0
KIAP	46.06	7.37	4.89	6.74	8.55
LBK Fiction	67.99	2.07	0.97	1.11	1.89
Difference	-21.93	5.30	3.92	5.63	6.66

Table 2: Results from coverage tests of five different lists in two test corpora.

As table 2 shows, the Random list has a high coverage in both corpora, but is highest in the LBK Fiction corpus, which is to be expected. Coverage is counted as the percentage of the number of lemmas in the corpus that occur in the list. All our experiments with the AKA lists show that to some extent the degree of coverage in KIAP correlates with the degree of coverage in LBK Fiction. A high coverage in both means that the AKA list has a high number of general (and not infrequent) words. Determining which list is the best is therefore not straight forward, judging from the coverage numbers alone. We therefore added a Difference measure, which shows the difference between the coverage for each list in the two corpora. However, this turned out not to be enough. A manual evaluation of the lists showed that those that had coverage of about 1.5 in LBK Fiction were too general.

In Table 2 there are two lists that satisfy the requirement of having less than 1.5 coverage: GM-2000 (0.97) and G&D-2.6_0.3_0.6_3.2 (1.11). Looking at their coverage in the academic KIAP test corpus, however, the G&D list is far better: 6.74 against 4.89. This is confirmed in the Difference as well: 5.63 against 3.92.

6. Towards a Final List

Recall, from Section 2, that the main aim of the academic word list is a list that can aid students in their struggle to understand and produce Norwegian academic texts. Looking again at Table 2, we see that the AKA list G&D-2.2_0.4_0.6_3.0 has as much as 8.55 coverage in the KIAP academic test corpus. This means that a lot of academic words have been lost compared with the G&D-2.6_0.3_0.6_3.2 list, which has 6.74 in KIAP. We therefore decided to manually merge the two lists and in the process remove those that were judged to be most

general. A lexicographer assisted in this latter task. The new, resulting 750-word list has coverage of 8.1 in KIAP and 1.3 in LBK-fiction.

The numbers are comparable to those claimed by Nation (2001), and to the 652 Swedish academic word list which has a coverage of 8.7 % in an academic corpus (Carlund et. al 2012). Gardner & Davies (2013:19) obtain 13.8 in coverage for their own list in the COCA academic corpus, compared with 7.2 for Coxhead (2000)'s list. It is interesting that the coverage is so similar for so many lists across languages. Gardner & Davies (2013:19) have a higher coverage, but their list also has a much higher coverage in the other genres: 8 % in a newspaper corpus and 3.4 % in a corpus of fiction. Their list therefore obviously contains many more words from the general vocabulary. There is no clear rule for how to judge which academic list is best, it all depends on the purpose of the list.

Our final list is presented and downloadable from a website, see Figure 2.

The web page contains the 750 words, sortable by frequency or alphabetic order. It also presents the part of speech for each word and examples from the DUO corpus. The words in Figure 2 are: *kunnskap* 'knowledge', *samfunn* 'society', *undersøkelse* 'investigation', *betydning* 'importance', *erfaring* 'experience', *forståelse* 'understanding', *nevne* 'mention', *tema* 'theme', *imidlertid* 'however', *analyse* 'analysis'.


The web page further offers definitions from the main standard dictionary *Bokmålsordboka*, and even the opportunity to paste in a text and check its academic vocabulary. It is to be hoped that this site will be used by students and teachers alike.

A Akademisk ordliste – bokmål Lista Er teksten akademisk? Om lista About

Søk i ordlista

Listenr	Ord	Ordklasse	Eksempler
21	kunnskap	substantiv	klikk for eksempler
22	samfunn	substantiv	klikk for eksempler
23	undersøkelse	substantiv	klikk for eksempler
24	betydning	substantiv	klikk for eksempler
25	erfaring	substantiv	klikk for eksempler
26	forståelse	substantiv	klikk for eksempler
27	nevne	verb	klikk for eksempler
28	tema	substantiv	klikk for eksempler
29	imidlertid	adverb	klikk for eksempler
30	analyse	substantiv	klikk for eksempler

TILBAKE NESTE START SLUTT


UiO : Institutt for lingvistiske og nordiske studier
 Det humanistiske fakultet

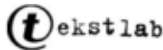


Figure 2: The web site for the Norwegian AKA list.

7. Conclusion

We have presented the development of a Norwegian Academic Wordlist (AKA list) for the Norwegian Bokmål variety. To identify specific academic vocabulary we developed a 100-million-word academic corpus based on the University of Oslo archive of digital publications. Other corpora were used for testing and for developing general word lists. We tried two different methods, those of Carlund et al. (2012) and Gardner & Davies (2013). The lists were compared measuring coverage in two different corpora. While it is good to obtain a high number for coverage in an academic corpus, this is also followed by a high number for general texts. We therefore experimented with the lists that we had, and inspected the intermediate results manually. Our final list has a coverage of 8.1 % in an academic test corpus, and 1.3 in a fiction corpus, which is comparable to the results of Coxhead (2000) and Carlund et al. (2012). While Gardner & Davies's list resulted in 13.8 % coverage, they also got very high coverage for general words: 8 % in a newspaper corpus and 3.4 % in a fiction corpus. Ultimately, it is the need of the users that decide which degree of coverage is acceptable. The resulting Norwegian Bokmål AKA list is presented on a web site, where the words can be inspected in different ways, and freely downloaded.

8. Acknowledgements

We would like to thank the University of Oslo Library for giving us access to their digital archive of publications. We would also like to mention that the KIAP corpus was accessed through the Clarin infrastructure Clarino. Professor Ruth Vatvedt Fjeld was central in the lexicographical evaluations, for which we are grateful. We would also like to thank our Scandinavian colleagues in the Nordplus language supported research network LUNAS – Language Use in Nordic Academic Settings. Our project was financed by the Department of Linguistics and Scandinavian Studies, University of Oslo.

9. Bibliographical References

- Carlund, Carina, Jansson, Håkan, Johansson Kokkinakis, Sofie, Prentice, Julia & Judy Ribbeck (2012). An academic word list for Swedish - a support for language learners in higher education In: Proceedings of the SLTC 2012 workshop on NLP for CALL, Lund, 25th October, 2012.
- Coxhead, A. (2000): A new academic word list. I: TESOL Quarterly, 34:2, 213–238.
- Gardner, D. & M. Davies (2013): A New Academic Vocabulary List. In: Applied Linguistics 4.
- Johannessen, Janne Bondi; Hagen, Kristin; Lylum,

- André; & Anders Nøklestad. (2012). OBT+stat: A combined rule-based and statistical tagger. In Andersen, Gisle (ed.). Exploring Newspaper Language: Using the web to create and investigate a large corpus of modern Norwegian, s. 51–66.
- Julliard, A. & E. Chang-Rodriguez. (1964). Frequency Dictionary of Spanish Words. The Hague Mouton.
- Lynum, Andre; Hagen, Kristin; Johannessen, Janne Bondi; & Anders Nøklestad. (2011). OBT+Stat: Evaluation of a combined CG and statistical tagger. In NEALT Proceedings Series 2011 Vol. 14.
- Nation, Ian SP. (2001) Learning vocabulary in another language. Ernst Klett Sprachen.
- Scott, M. (1997). PC analysis of key words—and key key words. *System*, 25(2), 233-245.
- West, M., & M. P. West (Eds.). (1953). A general service list of English words: with semantic frequencies and a supplementary word-list for the writing of popular science and technology. Addison-Wesley Longman Limited.

10. Tools and Web Sites

- Academic wordlist Norwegian Bokmål (AKA list):
<http://www.tekstlab.uio.no:4000/>
- Bokmålsordboka: <http://bokmålsordboka.uio.no/>
- BNC (British National Corpus):
<http://www.natcorp.ox.ac.uk>
- COCA (The Corpus of Contemporary American English):
<http://corpus.byu.edu/coca/>
- DUO University of Oslo digital publications archive:
<https://www.duo.uio.no>
- KIAP Corpus: <http://kiap.uib.no/KIAPCorpus.htm>
- LBK Corpus:
<http://www.hf.uio.no/iln/tjenester/kunnskap/samlinger/bokmal/veiledningkorpus/>
- LUNAS – Language Use in Nordic Academic Settings:
<http://cip.ku.dk/english/research/network/lunas/>
- LäSBarT Corpus:
<http://spraakbanken.gu.se/eng/resource/lasbart>
- NoWac Norwegian Web as a Corpus:
<http://www.hf.uio.no/iln/om/organisasjon/tekstlab/prosjekter/nowac/>