On Compiling a Norwegian Academic Vocabulary List

Ruth Vatvedt Fjeld, Arash Saidi

Department of Linguistics and Scandinavian Studies, University of Oslo
e-mail: r.e.v.fjeld@iln.uio.no, arashsa@student.ifi.uio.no

Abstract

In the last few years, several dictionaries of academic vocabulary have been developed, based on automatic analysis of academic corpora. Our presentation documents the first attempt to create a Norwegian Academic Vocabulary list (NAV) for the Norwegian Bokmål variety. This work is part of the Nordic project LUNAS (Language Use in Nordic Academic Settings) where the goal is to develop academic vocabulary lists and dictionaries for the Scandinavian languages. Part of the goal is to support Nordic language use in academic work instead of English, to maintain the Nordic languages also as a full-fledged academic language.

Keywords: academic vocabulary; academic words; academic corpus; statistical lemma selection; multiword expressions (MWEs); phraseology

1. Introduction

Nation (2001:11-13) has studied the coverage of different kinds of vocabulary in an academic corpus. He classifies words in academic texts into four types: High frequency words, which consist of function words and everyday content words, academic words, which are words that are common in different kinds of academic texts (e.g. Coxhead's 570 words), technical words, which are words closely related to the topic and subject of the relevant text, often called technical terms, and finally low frequency words, which makes «by far the biggest group of words» according to Nation. Low frequency words are not academic words, and not technical words for a particular subject. But they can be technical words from other subject areas, or proper nouns, or words that almost go into a high-frequency list, but are rare words in general language.

Nation finds that the first 1000 most frequent words in a text covers 71,4 % of the vocabulary, and only 4,7 % by the next 2000 most frequent words. Coxhead's Academic Word List of 570 words covers 10.0 % of the vocabulary.

Gardner & Davies (2013:8) state that academic core words are "those that appear in the vast majority of the various academic disciplines" in contrast to the 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". The purpose of our research is to test and evaluate a method for identifying and extracting academic core words as defined above, and to evaluate the pedagogical and lexicographical value of the lists.

2. Material: The Norwegian Academic Corpus DUO

Prior to our endeavours no general academic corpus existed for Norwegian. We therefore have assembled a set of academic texts and created an academic corpus of Norwegian language. To this end we used the Digital publications archive at the University of Oslo (DUO), which consists of master's theses, doctoral dissertations and scientific papers produced at the university. These documents have been downloaded in pdf format, converted to text, identified with respect to language (Norwegian Bokmål), and lemmatised with the Oslo-Bergen Tagger. There are 9689

documents in the corpus totalling approximately 310 million tokens. The documents come from all eight faculties of the University of Oslo (Humanities, Education, Medicine, Social and Economic studies, Mathematics and Natural sciences, Law, Theology and Odontology). The corpus is divided according to these and further after their respective departments. Consequently we have a rather fine grained subject classification of the texts.

3. Methodology

Our approach to compile the lists follows the method of Carlund et al. (2012), which is a modified version of the method utilised by Coxhead (2000, 2011). The method consists of four steps:

a. Keywords: A keyword is defined as a word that occurs with unusual frequency in a given text. Like Carlund et al (2012) we rank each word in the text according to its keywordness with a selection criteria to be a score of above 1.1. We use the web corpus NoWaC (Guevara) as the reference corpus.

b. Reduced frequency and range: For each word we divide the corpus in sets of intervals based on the frequency of the said word. We then count how many of these intervals the word appears in. This measure gives an indication of the extent to which a given word is spread out across the corpus. If a word has a high frequency in the corpus, but a low reduced frequency, we can conclude that this word belongs to the specialised vocabulary of a specific academic field. We removed words that did not have a reduced frequency of at least 15 per million words in each of the sub-corpora (i.e., departments). A mathematical description of this method can be found in Carlund et al (2012).

c. Removal of everyday words: Finally, using a stop list consisting of the most frequent words (based on a general language corpus), we remove words that have a high frequency in both academic and general corpora. These are words that have a low score on keywordness and a high reduced frequency, which means that they will not be excluded by the two steps above.

The words that are not removed in this three-step procedure are counted as candidates for an academic vocabulary list.

4. Preliminary Results

Results from analysis of the DUO-corpus with this three-step procedure, gave a list of candidates for an academic vocabulary list. A manual scrutiny of this automatic lemma selection showed few mishits in the list, and a comparison with the vocabulary of a general corpus (Leksikografisk bokmålskorpus (LBK) cf. Knutsen & Fjeld 2013) of 100 million running words, gave much higher frequency numbers in DUO than in LBK. We take this as an indication of these words' academic relevance. The automatically compiled lists made by means of different methods, were examined for the academic relevance, and to what degree these words could be counted as not common or frequent in non-academic texts as fiction or newspapers. An automatic way to figure that out is to match the frequency patterns of the words on the academic list with their frequency in general language.

A manual investigation of the 10 most frequent nouns, verbs, adverbs and adjectives on the list of potential academic words, showed that the frequency in DUO and LBK was widely different. The noun with number 1 frequency from DUO, kapittel (chapter), was number 21899 in LBK, the most frequent verb undertrykke (suppress, repress) was number 8 in DUO and 3292 in LBK, the

adjective teoretisk (theoretical) number 17 in DUO and 1975 in LBK, the adverb ibid. number 24 in DUO and 62151 in LBK. We take these significant differences as a clear indication that the vocabulary in our result list, is different than the frequency in general vocabulary. A human assessment of the candidates, also indicates that our method of word extraction has found good candidates for an academic vocabulary list.

5. Academic Phraseology

However, from our experience with student papers through many years, we know that more than single nouns and verbs, the words that are needed for writing good academic texts often are single function words like some adverbs and conjunctions, or multiword expressions, often with text structuring functions, or the function to modify or make reservations for the statements given.

In Norwegian academic language, several such expressions are adverbials that have the grammatical structure preposition+noun+preposition (PNP), like i forhold til (in relation to). Such phrases often consist of high frequency words, and hence they will be sorted out by the stoplist-methods when compiling academic words from a corpus. We therefore made a preliminary investigation of such phrases by means of statistical trigrams of presumptive academic texts categorized as non-fiction in LBK compared to such trigrams in fiction, like novels and short stories (cf. Fjeld & Saidi 2015).

Among the 500 most frequent trigrams in non-fiction, we found the 10 most frequent as below, but in the fiction subcorpus we did only find 6 such PNPs among the 500 most frequent words:

freq no	non-fiction	freq no	fiction
2	nouni forhold til (in relation to)	7	ved siden av (in addition to)
6	på grunn av (because of)	46	på grunn av (because of)
11	i løpet av (during)	59	i løpet av (during)
14	i forbindelse med (in connection with)	257	på vei til (headed for)
20	i tillegg til (in addition to)	473	i nærheten av (close to)
31	i form av (as)		
71	i henhold til (according to)		
78	i motsetning til (as opposed to)		
84	ved siden av (in addition to; nearby)		
85	på bakgrunn av (on the basis of)		

Table 1: Frequency table for non-fiction and fiction phrases.

This small investigation indicates that our hypothesis that academic texts differs from general language in several ways, not only in single content words is correct. Such words and phrases are more important to express or explain the real academic work, as described in Martin (1976:72).

6. Further work

The method we have used depends on the kind of stop lists used and the number of words they contain. The method described in Gardner & Davies (2013) does not use a stop list, so in the future we will compare our method with theirs. We will then test the validity of the lists by examining the coverage measure on academic texts in comparison to other kinds of text. We also want to work further with MWEs in academic language to include such phrases also in the Norwegian Academic Vocabulary List.

7. References

Coxhead, A. (2000). A new academic word list. In: *TESOL Quarterly*, 34:2, 213–238.

Carlund, Carina, Jansson, Håkan, Johansson Kokkinakis, Sofie, Prentice, Julia & Judy Ribeck (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.

Coxhead, A. (2011). The academic word list 10 years on: Research and teaching implications. In: *TESOL Quarterly*, 45:2, 355–362.

Fjeld, Ruth Vatvedt & Arash Saidi: Hvordan slippe inn i Platons hage? – Om kartlegging og dokumentasjon av norsk akademisk vokabular. In: Nordiske studier i leksikografi 13. *Rapport fra Nordisk konferanse om leksikografi i Norden, København* 2015 (in press)

Gardner, D. & M. Davies (2013). A New Academic Vocabulary List. In: *Applied Linguistics* 4.

Guevara, Emiliano Raul (2010). NoWaC: a large web-based corpus for Norwegian. In *Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop*, Association for Computational Linguistics, page 1 - 7.

Knudsen, Rune Lain & Ruth Vatvedt Fjeld (2013). LBK 2013: A balanced, annotated national corpus for Norwegian Bokmål. In: *Proceedings of the workshop on lexical semantic resources for NLP at NODALIDA 2013. NEALT Proceedings Series 19 / Linköping Electronic Conference Proceedings* 88, 12–20. <http://www.ep.liu.se/ecp/088/003/ecp1388003.pdf>

Martin, A. (1976). Teaching Academic Vocabulary to Foreign Graduate Students. In: *TESOL Quarterly* 10.

Nation, I.S.P. (2001) *Learning Vocabulary in Another Language.* Cambridge University Press