Developing Academic Word Lists for Swedish, Norwegian and Danish – a joint research project

Sofie Johansson Kokkinakis, Emma Sköldberg, Birgit Henriksen, Kari Kinn & Janne Bondi Johannessen

Keywords: *Academic Word List, Nordic Languages, Higher education, Language learning and teaching*.

Abstract

This paper reports on a joint multi-disciplinary Nordic project aimed at developing three new academic lexical resources based on corpora consisting of texts from Swedish, Norwegian and Danish academic settings. An academic word list exists for English, but no such lists exist for the Nordic languages. Such a list would be an important resource for both L1 and L2 students in their first years of study, a period when many students struggle to cope with the demands of academia. Moreover, the word lists would be of use to students and teachers at the higher levels of secondary education. An inventory of academic words and phrases would also be a useful tool for researchers of academic language use and for test developers. The paper outlines the initial stages of work on an academic word list for Swedish. Three potential research approaches have been explored: the translation of the English list, extracting academic words from existing corpora, and the compilation of parallel academic corpora where an academic word list is extracted from these. The paper will discuss the advantages and drawbacks of the different approaches and the benefits of carrying out a joint project involving several languages. The question of entry selection and the information categories of the dictionary entries and the interplay between the entries in the dictionaries and the corpora will also be briefly addressed.

1. Introduction

To be able to cope with the demands of academia, university students must not only master the technical vocabulary of their own field of study, they must also be able to understand and use a more general academic vocabulary which is frequent across a range of study areas. The second or foreign language user, who is studying in an English medium context, can draw on existing lists of academic words for English, e.g. *The University Word List* by Xue and Nation (1984). Best researched and most recent is Coxhead's *Academic Word List* (AWL, 2000; cf. Granger and Paquot 2010). These lists include words like *substitute*, *underlie* and *establish*, which are characterised by i) being relatively frequent in academic texts from a wide range of disciplines, ii) not being specifically connected with any one particular subject area, iii) being relatively infrequent in other types of texts, such as fiction, and iv) having a French, Latin or Greek origin (Wang Ming-Tzu and Nation 2004). English academic word lists have been used for developing academic language courses, and academic words drawn from these lists have been incorporated in various types of language tests (Nation 2001: 187–198). Recently, some work has also been started on developing similar academic word lists in other languages, e.g. Portuguese (Baptista et al. 2010; cf. Cobb and Horst 2004 for a French approach).

The principal aim of this paper is to present a new multi-disciplinary Nordic project (including disciplines such as linguistics, second language learning, lexicography and language technology) with the aim of developing three new academic lexical resources based on corpora consisting of texts from Swedish, Norwegian and Danish academic settings. Once finalised, the dictionaries and corpora will be publicly available. The Nordic collaboration in the project is fruitful from a research perspective, but collaboration is also an important prerequisite for the work to be economically sustainable as it concerns relatively small languages.

In the paper we outline, among other things, the principles and methods for the development of the dictionaries. We will discuss relevant existing language resources and

NLP tools in the three Nordic countries, and the development of new ones. Finally we will address the entry selection and, to some extent, the information categories of the dictionary entries and the interplay between the entries in the dictionaries and the corpora.

The Nordic academic word lists will not only be a useful tool for those students and teachers who are second language users of the three languages, they will also be an important resource for L1 students in their first years of study, a period when many students are struggling to cope with the demands of academia. Finally, it is our hope that the word lists will also be of use to students and teachers at the higher levels of secondary education.

This project must also be viewed from a language policy perspective. Recent documents from many Nordic universities have voiced a concern for the increasing role and use of English in academia at the expense of the national languages. For example, a study from the Language Council of Sweden has shown that 90% of all dissertations in Sweden are nowadays written in English (Salö 2010). Moreover, many MA courses are now conducted in English. At the Faculty of Humanities at the University of Oslo in Norway, the percentage of doctoral dissertations in English was nearly 60% in 2012, and even higher at the other faculties. Increased internationalisation in academia has the positive effect of wider dissemination of research findings and increased academic mobility, but the fact that teaching and research are increasingly conducted in English may lead to domain loss in the native languages within certain areas. Moreover, studies have indicated some negative effects on learning outcomes and study efficiency when lectures and interaction between Swedish students and teachers occur primarily in English (see Salö 2010: 8, 14–19). Therefore the academic word lists for the Nordic languages may be an important tool for establishing academic language competence development activities for staff and students at various levels of higher education. Note for instance the activities taking place at the language support centres at the University of Gothenburg <http://www.utbildning.gu.se/student/sprakhandledning> and the Centre for Internationalisation and Parallel Language Use (CIP) in Copenhagen <http://cip.ku.dk/english/>.

Before we expand on the project work, we will briefly present the academic word list developed by Coxhead. (Samples of the word lists can be found in Coxhead 2000, 2002; <http://www.victoria.ac.nz/lals/resources/academicwordlist/>).

2. A short presentation of Coxhead's *Academic Word List*

Coxhead compiled the AWL on the basis of a corpus of academic texts. The corpus included some 400 texts from a range of academic articles, textbooks and course books. The corpus itself was limited in size, comprising 3.5 million tokens from 4 subcorpora, each consisting of about 875,000 tokens. The texts were taken from four disciplines: arts, commerce, law and science. Each subcorpus included texts from seven different subject areas, which are presented in Table 1.

Table 1. Subject areas in Coxhead's Academic Corpus (Coxhead 2000: 220).

Arts	education, history, linguistics, philosophy, politics, psychology, sociology
Commerce	accounting, economics, finance, industrial, relations, management, marketing, public policy
Law	constitutional, criminal, family and medicolegal, international, pure commercial, quasi-commercial, rights and remedies

Science	biology, chemistry, computer science, geography, geology, mathematics, physics

Coxhead extracted lexical types of high frequency and range across the four discipline areas (see Coxhead 2002: 75). The 2,000 most frequent words from the *General Service List* (GSL, West 1953, a list of the most frequent words from general English which were considered useful for language learners) were then excluded from the word list. On the basis of this, Coxhead compiled a list of 570 word families. Each family consists of a headword and its closely related inflected and derived forms as defined by Level 6 of Bauer and Nation's (1993) scale. To take some examples from sublist 1, the word families for the headwords **approach**, **concept** and **constitute** consist of the following words:

(1) **approach**	approachable, approached, approaches, approaching, unapproachable
(2) **concept**	conception, concepts, conceptual, conceptualisation, conceptualise, conceptualised, conceptualises, conceptualising, conceptually
(3) **constitute**	constituencies, constituency, constituent, constituents, constituted, constitutes, constituting, constitution, constitutions, constitutional, constitutionally, constitutive, unconstitutional

The important principle behind the idea of a word family is, according to Bauer and Nation (1993:253), that once the base form or even a derived word is known, the recognition of other members of the family requires little or no extra effort. By presenting word families, Coxhead follows in the tradition of other authors of word lists for learners of English (e.g. West 1953, Xue and Nation 1984). She points out that this solution "is supported by evidence suggesting that word families are an important unit in the mental lexicon" (Coxhead 2000: 217–218).

The AWL is, as the label indicates, a compilation of academic words with no definitions or examples given and with no links to the corpora provided. It consists of 570 word families divided into ten sublists based on the frequency and dispersion of the word families included in each subcorpus. Sublist 1 consists of the most frequent word families, while sublist 2 consists of the second most frequent word families, etc. As already mentioned, in sublist 1 you find the nouns **approach** and **concept**, but also for example **factor**, **function**, **method, period, process** and **structure**. Among the headwords you also, apart from **constitute**, find verbs like **establish**, **estimate**, **indicate**, **involve**, **occur** and **require** and adjectives such as **evident**, **individual**, **significant** and **specific**. The sublists are composed of word families, the headword indicated with bold style followed by potential family members. There is, however, no information regarding pronunciation, inflection or use of the headwords or family members. The only information provided to the user is the most frequent word marked in italics.

According to Wang Ming-Tzu and Nation (2004: 292–293), the principal aim in making the list was to provide an explicitly described, feasible vocabulary learning goal for learners of academic language. The list could be used for direct learning and teaching and in the design of teaching materials. The range of the different headwords from the AWL in relation to the various discipline areas and subcorpora has, however, been questioned (Hyland and Tse 2007). Furthermore, as Paquot (2007) points out, the classification into word families, without information on frequency, is not particularly helpful as not all members of a word family are likely to be equally frequent in academic texts (cf. for example the headword **item** with two of

its word family members, the verb *itemise* and noun *itemisation*.) Consequently, all the family members are not so useful to the learners. Finally, the fact that the words are listed without additional information such as definitions, inflection properties or examples makes the list difficult to use for guidance to text reception and text production (cf. Tarp 2008; Svensén 2009).

It is also important to stress that the AWL is an inventory of individual academic words which does not include academic formulas typical of academic language (see e.g Simpson-Vlach and Ellis 2010). It is likely that a range of these formulas are important structuring devices for academic texts and academic discourse and therefore will be good candidates to include in an academic resource.

3. Methods of creating word lists

Regarding the present progress of the work on academic word lists for the three languages involved, work on a Swedish academic word list has been initiated. During the process, various methods have been considered. One of them is the potential use of a translation approach. The AWL would then be translated from English to each of the languages in this project. The AWL has many Nordic cognate counterparts which play an important role in Nordic academic texts (e.g. *analysis*, *indicate*, *category*; cf. Cobb and Horst 2004). The translation approach has revealed some problematic issues regarding homography and polysemy of words in the AWL (Sköldberg and Johansson Kokkinakis in press). In an attempt to automatically translate the AWL, Lexin (Lexicon for immigrants) – a freely available English-Swedish dictionary – has been used. 27% of the 570 headwords in the word list, for example *consist*, *distribute* and *evaluate*, have a Swedish equivalent which could be considered a potential candidate for a Swedish list. These translations need to be checked manually, but, as mentioned, they may constitute interesting candidates for a Swedish academic word list. The same issues will also be relevant for Danish and Norwegian.

It remains to be seen whether these words constitute a representative selection of Swedish academic words. We plan to use the same approach as Coxhead (2000), i.e. compiling an academic corpus and extracting an academic word list from the corpus. In order to be able to compare the various Nordic academic word lists to each other regarding frequency and range, it is an advantage that classification of data, methods of extraction and contents of the lists are similar (cf. Nesi 2002). This will be a prerequisite for comparative research studies across the three languages. There are, however, few available corpora of academic texts. We have compiled three different ones for Swedish, and a couple of corpora have been collected at the University of Oslo and the University of Bergen. To our knowledge, no such corresponding corpora for Danish are available.

Regarding the Swedish word list, several approaches were applied for corpus compilation.[1] As a result of these different approaches, different Swedish corpora of academic texts exist. The first corpus, consisting of approximately 800,000 tokens, was compiled from nine recently published linguistic dissertations from the University of Gothenburg. The second, consisting of 20 million tokens, is compiled (more randomly) using the tool WebBootCaT in Sketch Engine (Kilgarriff et al. 2004; Baroni et al. 2006). It includes more than 900 documents, primarily concerning subjects such as economics, education and informatics. The third corpus is a collection of documents from the field of arts that are published and available from a national academic on-line database. This corpus, of about 11 million tokens, comprises approximately 220 documents by more than 140 different authors. A majority of the texts in the three corpora have been been tokenised, lemmatised and pos-

tagged (using natural language processing tools in Språkbanken (<http://spraakbanken.gu.se>) at the University of Gothenburg) and loaded into the Sketch Engine (Jansson et al. 2012).

We seek lexical items of high frequency and distribution in the texts from the academic corpora. These words will, however, not be frequent in language of more colloquial style or be contained in a base vocabulary. Coxhead (2000) excluded words from the General Service List (West 1953) from the words extracted from the academic corpora. No corresponding GSL for Swedish has been compiled, so a corpus of Swedish novels will be used as a reference corpus of colloquial language. We intend to use the keyword function in Sketch Engine as a means of comparison.

For the Norwegian and Danish word lists, so far no resources have been made specifically for this project. In relation to the Norwegian list, we will consider what can be gained from Fløttum et al. (2006) and Fløttum (ed.) (2007). The University of Oslo has a system called DUO in which all Master's and many Ph.D. theses from all subjects have been stored electronically. We hope to be able to use some of these to build a corpus of academic texts in Norwegian. There are also some academic texts in the Norwegian Lexicographical Bokmål Corpus. Of special relevance are the categories teaching books, non-fiction, academic theses, reports, and legal documents (altogether 148 texts.) In addition, there are about six million words of legal language and government reports in the Oslo Corpus of Tagged Norwegian Texts, Bokmål. These corpora can be found at the Text Laboratory, University of Oslo. In addition there is a corpus of 150 Norwegian articles of linguistics, economics and medicine in Bergen, the KIAP corpus. Frequency lists from these corpora will be created at the Text Laboratory, using a combination of the corpus tool Glossa and manual linguistic and programming efforts. The Danish project plans to collaborate with Det Danske Sprog- og Litteraturselskab and the DANTERM Centre in establishing available academic corpora for the project.

4. Conclusion

The aim of this project is, as outlined above, to develop web-based academic word lists in Swedish, Norwegian and Danish. The work presented will first of all be carried out by postdoctoral researchers and PhD students from a range of disciplines such as linguistics, second language learning, lexicography, and language technology in the three countries under the guidance of a group of senior researchers from the countries involved.

We find an appropriate approach is the one adopted by Coxhead (2000) in the compilation of her AWL, i.e. the extraction of academic word families from a well-defined academic corpus from a representative range of subfields and disciplines. However, this approach must be adapted to the conditions found in the Nordic academic settings. The access to and range of texts will be different. Moreover, the circumstances found in the three Nordic countries may not be totally identical.

Furthermore, the content of the dictionaries must of course suit the structural characteristics of the Nordic languages. In developing the word lists we must also pay attention to the traditional content of monolingual learners' dictionaries intended for different dictionary use situations (reception, production etc.) and the results of user studies.

Finally, it is important to discuss which types of links should be made between the word lists themselves and the corpora, so that the users of the word lists will have access to examples and concordances showing various usages in different co- and contexts. These data-driven examples will also be central in relation to developing and exemplifying the definitions given.

A range of important but also very interesting questions remain to be answered. The corpora and the word lists can play a crucial role in guiding Nordic language learners in their independent study. Furthermore, the lists can guide teachers when setting vocabulary goals for language courses as well as course and material designers in developing learning activities. The corpora and the dictionaries also provide a useful basis for further cross-linguistic comparative research – from several perspectives – into the nature of contemporary Nordic academic language use in different settings and genres across the three languages. The resulting lexical resources will also support an increased use of Nordic languages instead of the predominant English influence in academic and scientific discourse.

Notes

[1] The work was partly financed by the project University of Gothenburg's Language Year. Emma Sköldberg has carried out the research as part of her position financed by The Knut and Alice Wallenberg Foundation.

References

A. Dictionaries

West, M. 1953. *A general service list of English words*. London. (GSL)

B. Other Literature

Baptista, J., N. Costa, J. Guerra, M. Zampieri, M. Cabral and N. Mamede 2010. 'P-AWL: Academic Word List for Portuguese.' In: *Computational Processing of the Portuguese Language, Lecture Notes in Computer Science, 2010*, Volume 6001/2010, 120–123.

Baroni, M., A. Kilgarriff, J. Pomikálek and P. Rychlý 2006. 'WebBootCaT: A Web Tool for Instant Corpora.' In: E. Corino et al. (eds.), *Proceedings XII Euralex International Congress. Atti del XII Congresso Internazionale di Lessicografia : Torino, 6-9 settembre 2006*. Alessandria: Edizioni dell'Orso, 123–131.

Bauer, L. and P. Nation 1993. 'Word Families.' *International Journal of Lexicography* 6/4: 253–279.

Cobb, T. and M. Horst 2004. 'Is there Room for an Academic Word List in French?' In: P. Bogaards and B. Laufer (eds.), *Vocabulary in a Second Language. Selection, Acquisition, and Testing*. Amsterdam: John Benjamins, 15–38.

Coxhead, A. 2000. 'A New Academic Word List.' *TESOL Quarterly* 34:2, 2000: 213–238.

Coxhead, A. 2002. 'The Academic Word List: A Corpus-based Word List for Academic Purposes.' In: B. Kettemann and G. Marko (eds.), *Teaching and Learning by Doing Corpus Analysis. Proceedings of the Fourth International Conference on Teaching and Language Corpora, Graz 19–24 July, 2000*. Amsterdam/New York: Rodopi, 73–89.

DUO Digital publications at the University of Oslo. 15 March 2012. http://www.duo.uio.no/englishindex.html.

Fløttum, K. (ed.) 2007. *Language and Discipline Perspectives on Academic Discourse.* Newcastle : Cambridge Scholars Publishing.

Fløttum, K., T. Dahl and T. Kinn. 2006. *Academic Voices – Across Languages and Disciplines*. Amsterdam: John Benjamins.

Granger, S. and M. Paquot 2010. 'The Louvain EAP Dictionary (LEAD).' In: A. Dykstra and T. Schoonheim (eds.), *Proceedings of the XIV Euralex International Congress, Leeuwarden 6–10 July 2010.* Ljouwert: Fryske Akademy / Afuk, 321–326.

Hyland, K. and P. Tse 2007. 'Is There an "Academic Vocabulary"'? *TESOL Quarterly* 41:2: 235–253.

Jansson, H., S. Johansson Kokkinakis, C. Ribeck and E. Sköldberg 2012. 'A Swedish Academic Word List: Methods and Data.' In this volume.

KIAP corpus. 15 March 2012. http://www.kiap.uib.no/KIAPCorpus.htm.

Kilgarriff, A., P. Rychlý, P. Smrz and D. Tugwell 2004. 'The Sketch Engine.' In: G. Williams and S. Vessier (eds.), *Proceedings of the Eleventh EURALEX International Congress EURALEX 2004 Lorient, France, July 6-10, 2004*. Lorient: Université de Bretagne-Sud, 105–116.

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

Norwegian Lexicographical Bokmål Corpus. 15 March 2012. http://www.hf.uio.no/iln/tjenester/kunnskap/sprak/korpus/skriftsprakskorpus/lbk/index.html.

Nesi, H. 2002. 'An English Spoken Academic Word List.' In: A. Braasch, and C. Povlsen (eds.), *Proceedings of the Tenth EURALEX International Congress, EURALEX 2002, Copenhagen, Denmark, August 13–17, 2002.* Copenhagen: *CST*, 351–357.

Oslo Corpus of Tagged Norwegian Texts. 15 March 2012. http://www.hf.uio.no/iln/tjenester/kunnskap/sprak/korpus/skriftsprakskorpus/oslo/index.html.

Paquot, M. 2007 'Towards a Productively-oriented Academic Word List.' In: J. Walinski, K. Kredens and S. Gozdz-Roszkowski (eds.), *Corpora and ICT in Language Studies. PALC 2005*. Frankfurt am Main: Peter Lang, 127–140.

Salö, L. 2010. *Engelska eller svenska? En kartläggning av språksituationen inom högre utbildning och forskning*. (Rapporter från Språkrådet 1.) Stockholm: Språkrådet.

Sköldberg E. and S. Johansson Kokkinakis (forthcoming). 'A och O om akademiska ord. Om framtagning av en svensk akademisk ordlista.' In: *Nordiska studier i lexikografi 12*. Lund: Nordiska föreningen för lexikografi.

Simpson-Vlach, R. and N. C. Ellis 2010. 'An Academic Formulas List: New Methods in Phraseology Research.' *Applied Linguistics* 31 (4): 487–512.

Svensén, B. 2009. *A Handbook of Lexicography. The Theory and Practice of Dictionary-Making*. Cambridge: Cambridge University Press.

Tarp, S. 2008. *Lexicography in the Borderland between Knowledge and Non-Knowledge. General Lexicographical Theory with Particular Focus on Learner's Lexicography*. (Lexicographica Series Mayor 134.) Tübingen: Max Niemeyer.

The Text Laboratory, University of Oslo. 15 March 2012. http://www.hf.uio.no/iln/english/about/organization/text-laboratory/.

Xue, G. and I. S. P. Nation 1984. 'A University Word List.' *Language Learning and Communication* 3, 2: 215–229.

Wang Ming-Tzu, K. and I. S. P. Nation 2004. 'Word Meaning in Academic English: Homography in the Academic Word List.' *Applied Linguistics* 25/3: 291–314.