A Swedish Academic Word List: Methods and Data

Håkan Jansson, Sofie Johansson Kokkinakis, Judy Ribeck & Emma Sköldberg

Keywords: *Swedish language*, *language learning and teaching*, *academic vocabulary, corpus-based dictionary, corpus compilation*.

Abstract

Academic language often presents a challenge to students, both language learners and native speakers. Therefore there is a need for educational language tools such as academic vocabulary resources. To date, resources developed have mainly focussed on learners of English; similar support is not yet available for Swedish. This paper reports on three different approaches to compiling a corpus of authentic academic text material used in academic settings. The purpose is to compose an empirical basis for the construction of a Swedish academic word list which can be used in language teaching. Because we have chosen to follow the method used for the creation of *The Academic Word List* (Coxhead 2000), the corpus content is crucial to the final content of our word list.

1. Introduction

Language in academic educational settings often presents a challenge to students, both language learners and native speakers. This has resulted in the development of academic vocabulary resources, so far mainly for learners of English. Most influential is *The Academic Word List* (Coxhead 2000, see section 2), which contains words believed crucial to higher education independent of study orientation, for instance *analyse*, *distribution* and *indicate*. Also, academic vocabulary is highlighted in some general learners' dictionaries of English. However for students of the Swedish language, similar support is not yet available (see Johansson Kokkinakis et al. 2012).

In this paper we present different approaches to compiling a freely available web-based corpus of Swedish academic texts. This corpus will serve as a basis for extracting an academic word list.

Today a great deal of higher education in Sweden is in English and almost 90% of all Ph.D. theses are written in English (Salö 2010). At the same time, The Language Council of Sweden, among others, expresses a wish for Swedish researches to be able to communicate their results in their mother tongue. To this end, our word list would be an important resource. It would also be interesting from a lexicological perspective as a mapping of the Swedish academic vocabulary, which to date is a scarcely researched field.

We would also like to stress the importance of using text corpora as a method for constructing a word list based on empirical evidence, rather than subjectively choosing the words to be included in the list. Also, this corpus-based approach ensures that contents are validated and that language samples are authentic.

2. Coxhead's *Academic Word List*

In the late 1990s, Coxhead presented her *Academic Word List* (AWL) for English. The entries in the AWL are word families, each of which is a stem plus all closely related affixed forms (Coxhead 2000). An example of a word family is: *contribute - contributed, contributes, contributing, contribution, contributions, contributor, contributors.* The AWL contains 570 word families frequently found in Coxhead's Academic Corpus, a corpus compiled for the purpose of the AWL. The word families are not among the 2,000 most frequently occurring English words, as described in *The General Service List* (West 1953).

The Academic Corpus consists of 3.5 million tokens. It contains 414 texts (mainly articles and textbooks) by more than 400 different authors. The data is spread equally across four disciplines: the arts, commerce, law and science. Each discipline is divided into seven subject areas. The arts contains subject areas such as education, history and psychology. To be included in the AWL, the members of a word family cumulatively had to occur at least 100 times in the entire corpus, ten times in each of the four disciplines and in 15 of the subject areas.

Since its release, the AWL has hugely influenced the curricula of English for academic purposes and English as a second/foreign language (Hyland and Tse 2007, Granger and Paquot 2009). Nevertheless, Coxhead's selection methods and presentation have been criticised. Like Hyland and Tse (2007), one can certainly question Coxhead's division into disciplines and subject areas. As Nesi (2002) points out, it would be favourable if the division were transferable across institutions to enable comparison of different academic corpora. We believe that the difference in the word list's coverage within different disciplines and the dominance of commerce words, reported by Coxhead (2000), have to do with the fact that commerce is more homogenous than for instance science.

Eldridge (2007) and Hyland and Tse (2007) also question the usability of the actual list – for reception and production, as well as the benefit of word families for learners at different proficiency levels. They call for sense descriptions in general and subject-specific senses in particular, as well as combinatorial properties in relation to the words. They argue that the members of a word family should rather be taught separately, since their collocational patterns tend to differ.

3. Towards a Swedish Academic Word List: some possible methods

The methods used for developing academic word lists have been diverse, ranging from observation of learners' difficulties to calculation of frequency in university textbooks (Nesi 2002: 352). In an effort to produce a Portuguese equivalent to the AWL, translation is used (Baptista et al. 2010). It is also feasible – and to some extent tempting – to simply translate the AWL into Swedish. In fact, both manual and automatic translation of the AWL have been tested and evaluated. Yet the results show that translation of academic vocabulary is not trivial (Johansson Kokkinakis et al. 2012, Sköldberg and Johansson Kokkinakis, in press).

In the following we describe three methods, each of which is comparable to Coxhead's method, with the aim of compiling a corpus which can form the basis of a Swedish academic word list.[1] In general we aim at a larger corpus of Swedish texts than Coxhead's 3.5 million word corpus. Another aim is to include a wide range of subjects as equal in size as possible. Preferably several texts genres and texts written by different authors would also be included.

3.1. A corpus of linguistic theses

The idea of the first method is to collect texts in our immediate vicinity. So far we have a corpus of about 800,000 words consisting of nine recently published linguistic theses from our own department at the University of Gothenburg. All the authors are known to be native speakers of Swedish, which is seen as an advantage, since differences between L1 and L2-speakers' use of academic language have been reported (Granger and Paquot 2009).

Just like Coxhead (2000), we have dispensed with bibliographic information, but also for example extensive quotations and language excerpts, for instance youth language, that do not serve our purpose. A random check tells us that this procedure can reduce the number of words in a thesis by more than 15%. The texts have been tokenised, lemmatised and pos-tagged and the corpus has been loaded into the Sketch Engine (Kilgarriff et al. 2004).

Manual processing of the texts has been time consuming. The applicability of this corpus is also limited, since it only contains linguistic texts.

3.2. A corpus compiled using WebBootCaT

The second method is based on collecting a corpus by means of the tool WebBootCaT in the Sketch Engine (see Baroni et al. 2006). The tool was used in its advanced mode, with 20 random medium-frequency seed words from the 114 million word Swedish corpus SwedishWaC that is included in the Sketch Engine. The URLs were specified to domains of Swedish universities and the document type to .pdf. White lists and black lists were used to ensure that the linguistic material was in Swedish; the specification of the black list was set to allow brief quotations in English. Keywords were taken from the collected documents and then used to refine the seed words for subsequent searches. This procedure quickly resulted in a 20 million token corpus made up of 912 different documents.

However, getting more detailed information about the data required the documents to be manually assessed with respect to subject and document type. The assessment showed that the data was not uniformly distributed across university disciplines (see Fig. 1).

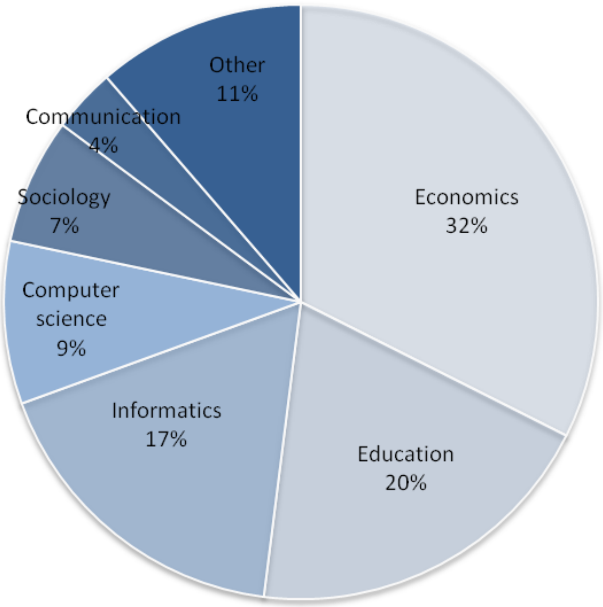

Figure 1. The most represented university subjects in a corpus collected using WebBootCaT.

In fact only six university subjects, such as economics, education and informatics, were represented to a relatively satisfactory extent. On the other hand, almost 90% of all the documents belonged to these six subjects. The few remaining documents were spread across at least 15 other subjects, none of which were sufficiently represented. Consequently, these subjects did not qualify for being subcorpora in an academic corpus.

It was unfortunate that the subjects included were few and diverse, because we would like to see all university subjects equally represented in the corpus. The limited number of documents collected from, for example, science can in all probability be linked to the available selection. According to Salö (2010), in Sweden 94% of the theses presented within the field of science are written in English.

3.3. A corpus with data from SwePub

The third method involves collecting published documents from a national academic on-line database, SwePub <http://swepub.kb.se/>, kept by the National Library of Sweden. All the documents have been catalogued in compliance with the guidelines set by the Swedish National Agency for Higher Education, which in turn are based on the OECD classification Field of Science and Technology (OECD 2007; cf. Nesi 2002 above).

At this point we decided to compile the corpus using theses and other academic publications from the arts, which is the most widely represented field in Swedish (see Salö 2010). The subjects chosen were *ethnology*, *history*, *linguistics*, *literature*, *philosophy* and *religious studies*.

The corpus comprises approximately 220 documents by more than 140 authors and contains roughly 11 million tokens (punctuation marks excluded). It has been divided into subcorpora with regard to the already mentioned subjects, as well as the document types *Ph.D. theses*, *Articles*, and *Other*. SwePub allows searches with the above specifications, so the corpus compilation was uncomplicated, although each document had to be downloaded manually.

Table 1. Number of words in a corpus of documents from SwePub.

	Ph.D. theses	***Articles***	***Other***	**Total**
Ethnology	1,210,735	69,047	168,712	**1,448,494**
History	2,119,048	93,721	95,312	**2,308,081**
Literature	1,753,839	205,482	26,616	**1,985,937**
Linguistics	1,544,166	156,921	228,058	**1,929,145**
Philosophy	454,266	48,157	140,892	**643,315**
Religious studies	2,282,125	48,794	288,615	**2 619,534**
Total	**9,364,179**	**622,122**	**948,205**	**10,934,506**

Table 1 shows the distribution of words in the corpus. As can be seen, the subcorpora vary in size. More specifically, philosophy is considerably smaller and ethnology somewhat smaller than the other subjects, but this reflects the total amounts of documents in the database. Like the corpus of linguistic theses, this corpus has been tokenised, lemmatised and pos-tagged and loaded into the Sketch Engine. The corpus is also freely accessible through Språkbanken <http://spraakbanken.gu.se/korp/#corpus=sweachum>.

4. Discussion

This paper discusses three methods for compiling a corpus, which can form the basis of a Swedish academic word list.

First, we compiled a corpus of 800,000 tokens consisting of easily accessed electronic texts, namely nine linguistic theses from the Department of Swedish at the University of

Gothenburg. We manually removed sections not comprising academic language, for instance language excerpts. However, this corpus contains far too few texts to be sufficiently representative. Furthermore, it is limited in terms of subject matter. To sum up, at present this approach cannot be used for the purposes required until similar subcorpora from other university subjects have been added.

Second, we compiled a corpus of 20 million tokens from texts on the Internet using the tool WebBootCaT. One advantage of this method was a large and quickly created corpus. One disadvantage was that more time was needed to structure the contents, which is absolutely essential for the intended usage. Another downside is the difficulty in gathering documents from several different subjects solely by way of seed words. The subjects represented by most available Swedish documents tend to be greatly overrepresented. In itself, the automatic WebBootCaT method cannot serve as the only method for compiling a balanced academic corpus. Still, the collected and classified documents can naturally be included in a manually compiled corpus, if there is a need for the subjects at hand.

Last, we compiled a corpus of 11 million tokens, consisting of published documents listed in SwePub, a national database for academic texts sorted by subject. The corpus contains documents from six subjects from the field of arts. One advantage is that the division into academic fields is OECD-based, which implies a systematic approach. Naturally, it would be desirable to be able to extend the corpus with additional disciplines. This, however, would entail great limitations to the size of the subcorpora and thus the corpus as a whole; the subcorpora should be of equal size and the selection of texts is very restricted in some academic fields.

As already mentioned, this paper reports on three different approaches on compiling a corpus of authentic academic text material used in academic settings. The purpose is to compose an empirical basis for the construction of a Swedish academic word list, which can be used in language teaching. We have chosen to follow the method by Coxhead and therefore the content of the corpus is crucial for the final content of our word list. This is a work in progress and the methods discussed will be further weighted, evaluated and combined to decide the best possible starting point for a valid and reliable Swedish academic word list.

Note

[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: Longman, Green. (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.' *Computational Processing of the Portuguese Language, Lecture Notes in Computer Science* 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 of the XII Euralex International Congress*. Torino: Alessandria, 123–131.

Coxhead, A. 2000. ‘A New Academic Word List.’ *TESOL Quarterly* 34.2: 213–238.

Eldridge, J. 2007. ‘“No, There Isn’t an ‘Academic Vocabulary,’ But …” A Reader Responds to K. Hyland and P. Tse’s ‘Is there an ‘Academic Vocabulary?’’ *TESOL Quarterly* 42.1: 109–113.

Granger, S. and M. Paquot 2009. ‘Lexical Verbs in Academic Discourse: A Corpus-driven Study of Learner Use.’ In: Charles, M. et al. (eds.), *Academic Writing: At the interface of corpus and discourse.* London/New York: Continuum, 193–214.

Hyland, K. and P. Tse, 2007. ‘Is There an “Academic Vocabulary”?’ *TESOL Quarterly* 41.2: 235–253.

Johansson Kokkinakis, S., E. Sköldberg, B. Henriksen, K. Kinn & J. Bondi Johannessen 2012. ‘Developing Academic Word Lists for Swedish, Norwegian and Danish – a Joint Research Project.’ In this volume.

Kilgarriff, A., P. Rychlý, P. Smrz and D. Tugwell 2004. ‘The Sketch Engine.’ In Williams, G. and S. Vessier (eds.), *Proceedings of the Eleventh EURALEX International Congress, Lorient, 2004*. Lorient, 105–116.

OECD. Organisation for Economic Co-operation and Development, 2007: Working Party of National Experts on Science and Technology Indicators: Revised Field of Science and Technology (FOS) Classification in the Frascati Manual. 15 March 2012. http://www.oecd.org/dataoecd/36/44/38235147.pdf.

Nesi, H. 2002. ‘An English Spoken Academic Word List.’ In Braasch, A. and C. Povlsen (eds.), *Proceedings of the Tenth Euralex International Congress*, *Copenhagen 2002* (vol. 1). Copenhagen, 351–357.

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

Sköldberg, E. and S. Johansson Kokkinakis (in press). ‘A och O om akademiska ord. Om framtagning av en svensk akademisk ordlista.’ In *Nordiska studier i lexikografi* 12. Lund.