An academic word list for Swedish

- a support for language learners in higher education

Carina Carlund
carina.carlund@svenska.gu.se
Håkan Jansson
hakan.jansson@svenska.gu.se
Sofie Johansson Kokkinakis
sofie@svenska.gu.se
Julia Prentice
julia.prentice@svenska.gu.se
Judy Ribeck
judy.ribeck@svenska.gu.se

All authors at the Department of Swedish
University of Gothenburg, Sweden

Abstract

The paper describes the ongoing development of compiling and introducing a Swedish academic word list (SAWL), inter alia intended to be used as a lexical resource in CALL-applications in relation to higher academic studies. When it comes to language acquisition, resources like these play an important part in instructed language learning. So far, no such resource exists for Swedish. The format of SAWL has been elaborated in collaboration with the Language Support Service at the University of Gothenburg. SAWL is compiled with methods from corpus linguistics inspired by research on English academic words (Coxhead 2002). Our work includes collection and syntactic annotation of learner corpora of Swedish academic texts from a wide range of university subjects within the Faculty of Arts. The corpora are freely accessible through Språkbanken. SAWL are designed with university students and language learners with Swedish or other linguistic backgrounds in mind. The word list and the corpora can be used for studies of one's own or in classroom situations, as well as forming a component of computer computer-based language assessment and CALL-related application platforms.

1 Introduction

The language in academic studies and in teaching is often a challenge for both L1 students without an academic background and L2 students. In order to meet the language demands, university students must not only master a subject's specific vocabulary, but also be able to understand and use a more general academic vocabulary, which is common within a range of study areas. To meet the students' need for knowledge of this type of vocabulary a number of English academic word lists have been developed. Our aim is to compile and offer a similar resource in the Swedish academic context.

An academic word and phrase list would serve as a valuable resource for L2 students in particular, but also for L1 students during their first year of university studies, a period during which many students struggle to meet the demands set by their academic studies, not least linguistically. In order to master both written and spoken academic language use, one has to be able to understand and use conventionalized formulaic expressions that are typical for academic discourse. Hence, in addition to a list of individual academic words, L2 students and students who are lacking experience of academic studies can be expected to have use for a resource that lists and describes multi-word expressions that are relevant for Swedish academic language (c.f. Ellis et al. 2008:379).

The Language Support Service at the University of Gothenburg had conducted a small user study of words in academic text and further user studies are planned. It is thought that the word and phrase list will be used in the language tutors' work, by other course teachers and by the students themselves. Today the development is towards computer based applications in the teaching of language and an academic word and phrase list is a resource that is suitable for CALL.

Carina Carlund, Håkan Jansson, Sofie Johansson Kokkinakis, Julia Prentice and Judy Ribeck 2012. An academic word list for Swedish – a support for language learners in higher education. *Proceedings of the SLTC 2012 workshop on NLP for CALL.* Linköping Electronic Conference Proceedings 80: 20–27.

The project of compiling a Swedish academic word and phrase list, which is also part of a wider Nordic collaboration, must also been seen from a language perspective. New documents from many Nordic universities have expressed concern about the increased use of English within academia to the detriment of the national language. For example, a study from the language council in Sweden demonstrated that 20% of all Swedish theses are now written in English (Salö 2010).

Increased internationalization in the academic world has the positive effect of increasing dissemination of research results and has increased academic mobility, but the fact that teaching and research more and more are conducted in English can lead to domain loss of the native language in certain areas. In addition, studies have pointed out a number of negative effects on study results when lectures and the interaction between Swedish students and teachers are mainly conducted in English (see Salö 2010: 8, 14-19).

2 Previous research

There has been a few attempts on the creation of academic vocabulary resources, so far mainly for learners of English but also for Portuguese (Baptista et al. 2010) and French (Cobb and Horst 2004). In this paper we describe the English one, being the best documented. The Academic Word List (Coxhead 2000), contains words believed crucial to higher education independent of study orientation, for instance *analyze*, *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).

2.1 The Academic Word List for English

[In the late 1990s, Coxhead presented her Academic Word List (AWL) for English. She believed in her approach that the content of an academic word list should be based on relevant principles within corpus linguistics. Therefore Coxhead compiled a corpus of academic texts to be able to extract the word list from.

The Academic Corpus consists of 3.5 million tokens. It contains 414 texts (mainly articles and text books) 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 (see table 1).

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

Table 1. Subject areas in the four AWL disciplines (Coxhead 2000:220).

The arts discipline contains subject areas such as education, history and psychology. To be included in the AWL, the members of a word family (West 1953; Bauer and Nation 1993) 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. 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. The word families are not among the 2,000 most frequently occurring English words, as described in The General Service List (West 1953). By using the concept of word families Coxhead concur in the tradition of previous creators of vocabulary lists for language learners (cf. West 1953, Xue and Nation 1984). Her motivation for this choice is that the use of word families "is supported by evidence suggesting that word families are an important unit in the mental lexicon" (Coxhead 2000:217f.).

As the name indicates, the AWL is a plain word list. It consists of word families, graphically indicated with an initial head word followed by family members – in the case there are any. There is however no information on the head words' or the family members' pronunciation, grammatical paradigms, meaning or collocational properties. The fact that there is so little information included in the list limits its use in academic settings as well as its use for lexicographic purposes. Advice for language learners on how to use the list is described at:

<http://www.victoria.ac.nz/lals/resources/academicwordlist/>.

Criticism

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 favorable 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 Resources and Method

Building on previous work on academic word lists, as presented above, there would be two main routes for this project to pursue: One could either simply translate Coxheads English list into Swedish or one could compile a corpus of Swedish Academic texts.

3.1 Translation of Coxheads AWL?

The translation path has been followed by a similar Portuguese project (P-AWL, Baptista et al, 2010), and also by other similar projects. Thus a Finnish WordNet has been produced, applying translations techniques to Princton WordNet (Lindén and Carlsson 2010) and a Norwegian LEXIN learners dictionary has been made based on the translation of the corresponding Swedish dictionary, (Bjørneset 2001). There are however some limitations connected to the translation method.

Martola (2011) lists some of the shortcomings of the Finnish WordNet, which are tied to its translation from English. Apart from the culturally specific semantic problems pointed out by Martola, there are also issues that are of a more lexical/morphological nature, which are partly connected to Coxheads notion of word families. These problems came to light when 60 headwords from sublist 1 of the AWL were compared to their Swedish translation equivalents in the dictionary *Norstedts stora engelsk-svenska ordbok* (2000).

Only a few of the words, e.g. *percent*, are easy to translate. More than a third of the words e.g. *contact* and *issue* are homographs and most words are polysemous. The English word families will inevitably be split up in a translation. For a further discussion of the translation method and some of its issues, e.g. the problems with the implications of the notion of word families, c.f. Sköldberg and Johansson Kokkinakis (2012).

3.2 Corpus collection

The translation option was subsequently discarded. Instead a decision to aim for a Swedish corpus of academic texts was taken. After finishing some pilot studies, designed to evaluate the effectiveness of different corpus compilation methods, reported on in Jansson et al. (2012), it was decided to compile a corpus from documents published in the Swedish national academic online database, SwePub <http://swepub.kb.se/>, kept by the National Library of Sweden.

An advantage with the use of that particular source is that 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. Organisation for Economic Co-operation and Development 2007). This foundation of our corpus in an official typology of Academic subjects provides an unbiased text subjects division and facilitates an easy comparison between countries, since it falls back on an OECD standard. As noted above, Nesi (2002) stresses that more uniform corpus subdivisions across different languages and groups would enable comparison of different academic corpora.

It should be noted that the subject of one entire subcorpus of Coxheads e.g. *commerce*, compares to OECDs *business and management* which is a secondary subdivision of the field *social sciences* in OECD typology, Coxhead (2002:75), OECD (2007).

3.3 The Arts corpus

Since the use of the Swedish language is not evenly spread over the different fields of science, we decided to start with a 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. The SwePub database allows searches with the above specifications, so the corpus compilation was uncomplicated, although each document had to be downloaded manually.

	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 2. Subjects and text types in the Arts corpus

Table 2 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 SwePub-database.

The texts were first cleaned from markup and code by uploading them into the Sketch Engine (for ref. see Kilgarriff et al., 2004). Then they were downloaded and subsequently tokenised, lemmatised and pos-tagged at Språkbanken.

3.4 Word selection

The principle for word selection for the list is based on the aim of finding an academic-specific vocabulary that is common for all subjects at the university, but not part of the everyday language.

As pointed out by Savický and Hlaváčová (2002), there is no formal definition of the intuitive notion of "commonness" when trying to rank words of the language. Most often, absolute or relative frequency of words in a corpus has come to denote commonness. This however is far from an optimal measure.

To obtain a more objective measure of word commonness, one has to look not only at frequency, but also at the distribution of that frequency. This is what is done by means of different types of *corrected frequencies* (Savický and Hlaváčová 2002).

Reduced frequency

The sort of corrected frequency we applied is called *reduced frequency*, RF[1] (Hlaváčová 2000; Savický and Hlaváčová 2002) and is calculated as follows:

Let *f(x)* be the frequency of word x in a corpus consisting of *N* tokens. Then divide positions of the whole corpus into *f(x)* intervals $< i , j >$. For $n = 1 \ldots f(x)$, the *n*:th interval is:

$$< \ [(n-1)N/f(x) +1] \ , \ [nN/f(x)] \ >$$

Let *Fx* be the partial frequency of *x* as:

Fx(n) = 1, if x occurs in the *n*:th interval
Fx(n) = 0, otherwise

RF(x) is then simply the sum of all partial frequencies for *x*:

$$RF(x) = \sum_{n=1}^{n=f(x)} Fx(n)$$

RF ensures the frequencies to be spread across the corpus without requiring the corpus to be divided into sub corpora according to for example genres or text types. This is a great advantage to other measures of dispersion, since "any trial of text annotation brings plenty of problems, which are difficult, if not even impossible to resolve… Moreover there is no strict border between genres…" (Savický and Hlaváčová (2002:216f.).

The RF for evenly distributed words is closer to their absolute frequency, and the RF for unevenly distributed words is smaller than their absolute frequency.

Keywords

To automatically identify domain-specific vocabulary, we ranked the lemmas according to keywordness (Scott 1997). The reference corpus was set to a 2.5-millon token collection of novels from Nordstedts, available through Korp at Språkbanken. The first selection criterion we

[1] After conducting some tests on our material, we decided not to use the Average Reduced Frequency described in Savický and Hlaváčová (2002) and Hlaváčová (2006). The results showed that RF was sufficient, since the values of RF and ARF hardly differed.

applied was for a lemma to score above 1.1 in keywordness.

Range
The second selection criterion was a requirement for the lemmas to have a relative RF of at least 15 per million tokens in each of the university subjects. By applying this demand of range, we increased the remedy for the "burstiness" problem (Kilgarriff 2009), which still was salient in our preliminary list. Moreover, we wanted to be sure that the words really were common to all subjects included.

Some examples of lemmas ruled out at this stage were: *präst* 'priest', *världskrig* 'world war', *sexualitet* 'sexuality', *kung* 'king', *författarskap* 'authorship', *medeltid* 'Middle Ages', *lagstiftning* 'legislation', *ordbok* 'dictionary' and *syntaktisk* 'syntactic'.

Filtering out non-everyday words
The third selection criterion was that the lemmas should not be part of the most frequent words of everyday Swedish. The filtering was done by removing all lemmas that belonged to the 1000 most frequent words of the 1.1-million token corpus LäSBarT available through Korp at Språkbanken. This corpus contains children's books and other easily read texts.

Words ruled out at this stage were for instance: *svensk* 'Swedish', *exempel* 'example', *språk* 'language' and *istället* 'instead'.

Manual processing
The final step was to manually clean the list from unwanted noise, such as abbreviations like *s.* 'p.', *t.ex.* 'e.g.', *jfr* 'cf./cp.' and *eds.*, numerals and text-structuring tokens as *ii.*.

We also brought some entries together that were tagged as different parts-of-speech [2], although according to modern lexicographic tradition belong to the same entry. As an example, words tagged as both adjectives and adverbs, e.g. *speciell* 'special', only appears as an adjective in the final list.

4 The resulting list

Our methodology for identification of academic words has resulted in a word list of 750 entries.

4.1 Entries

The 10 topmost entries of the list according to keywordness are: *dock* 'however', *relation* 'relation', *samt* 'and', *studie* 'study', s*ocial* 'social, public', *begrepp* 'concept', *form* 'form', *betydelse* 'meaning, importance', *analys* 'analysis' and *utifrån* 'on the basis of'.

We regard the lack of information about the words in the AWL to be a drawback. The entries in our list are annotated with: [3]

1. part of speech
2. inflectional forms
3. meaning
4. one (or more) editorial examples based on instances in the corpus
5. English translations.

To exemplify what the entries look like, we can look at the word *innebära* (imply, mean).

> **innebära** (verb) *innebar*, *inneburit*; *innebär* • betyder, medför. *Vårdnadsansvaret innebär både rättigheter och skyldigheter för dig som förälder; Romerskt medborgarskap innebar en mängd friheter och privilegier*. 'imply, mean'.

As far as the meanings are concerned, all the meanings given in *Lexins svenska lexikon* (2011) are included, even the ones that may not be that common in the academic texts. This approach was chosen since not all instances of this dilemma were entirely intuitively obvious.

The examples should function as an aid to the information about meaning. The intention is that they should be illustrative of one of the given meanings – preferably the one most common in the corpus. To facilitate for the users, the examples are editorial, which means that they are based on authentic occurrences in the corpus, but depicted with less or simplified context when needed. In the online version of the list, the user can easily follow a link to the corpus and look at actual concordances.

[2] Pos-tagging was made by means of the open source hunpos-tagger, which implements the TnT-tagger. The tagger is trained on data from SUC 2.0 from which the pos-tags derive.

[3] So far, this work has been carried out for the first 100 entries by a lexicographer. The information about part of speech, inflection and meaning are drawn from the recently revised 4th edition of *Lexins svenska lexikon* (2011) supplied by Språkrådet. The English translations are taken from *Lexins svensk-engelska lexikon* supplied by Språkbanken.

4.2 Coverage

With regard to a previous categorization of word types, Nation (2001) concludes that the vocabulary of academic texts consists of 80% of the most common and frequent words, 8-10% general academic words and 5% subject specific and technical words.

The 750 words of our list cover on average 8.7% of our Arts corpus (10.1% linguistics, 7.9% history, 8.1% ethnology, 7.9% literature, 10.4% philosophy and 9.1% religious studies). This can be compared with the 10.0% coverage of the AWL reported by Coxhead (2000). We believe the smaller coverage of our list can be explained by at least three factors.

First and foremost, we apply much more rigorous selection criteria. The words of the AWL are chosen as a consequence of frequency and range alone, while we also require certain keywordness in relation to a reference corpus, as well as considering the distribution of the frequencies (dispersion). We strongly believe that this approach will assure a high precision of academic vocabulary. Besides that, total recall was never our goal. Most important for us was to identify a crucial vocabulary for academic achievements, in that knowledge of the words would help students in their academic studies.

Second, the entries of the AWL are word families (see 2.1), while we have lemmas. Word families may contain lemmas from different parts of speech as well as affixated word forms, e.g. [*available, availability, unavailable*]. Since the selection procedure for the items of the AWL adds the frequencies of all the members of a word family, not all members alone need to fulfill the requirement for inclusion. Still, these "additional" members contribute to the overall coverage of the AWL.

Third, academic texts written in Swedish contain a non negligible amount of non Swedish language, for example in citations and summaries. Since we only included Swedish words in the list, foreign language in the corpus was never going to be covered.

5 Conclusions and accessibility

This paper describes the use for and the creation of an academic word list for Swedish. The method describes an approach where a list of 750 lexical items is extracted from a compiled corpus of Swedish academic texts publically available through Språkbanken. The overall coverage of the word list is 8.7% of the corpus.

The word list is available, both from Språkbanken and as a freely downloadable lexical recource – *En svensk akademisk ordlista*, version 1.0, <http://spraakbanken.gu.se/ao/>.

The list is shown online and is downloadable in two formats. On the one hand, there is a listing of all the 750 headwords, which can be viewed in alphabetical order or according to keywordness. On the other hand, there is the fully annotated top-100 list of words according to keywordness.

6 Future research and applications

The described lexical resource, SAWL, is intended to be used in language learning both individually and in academic class room settings.

6.1 Research

The next immediate step will be an evaluation process of the usefulness of the extracted lexical items, in collaboration with the University of Gothenburg language support service.

As an extension to SAWL, inspired by the research carried out by Ellis et al (2008), and Simpson-Vlach and Ellis (2010) that resulted in an Academic Formula List (AFL) for English, we are also aiming to use various methods within the fields of i.e. language technology, corpus linguistics and psycholinguistics, to develop a list of conventionalized multi-word expressions for Swedish academic language. As Ellis et al. point out, it has been established in relatively recent research "that highly frequent formulaic expressions are not only salient but also functionally significant: Cognitive science demonstrates that knowledge of these formulas is crucial for fluent processing" (Ellis et al. 2008:379).

In addition to the research questions mentioned above, the next step in extending the corpus in the subject areas social sciences and natural sciences. The latter being more difficult since English is used more often in those subjects.

6.2 Applications

Regarding how to implement SAWL in computer-based applications, the aim is twofold; one goal is to use it as a validated and reliable lexical resource in language assessment platforms similar to those implemented in the testing part of the "Complete Lexical Tutor" for assessment of English general and language specific vocabulary tests <http://www.lextutor.ca> and in the testing of Swedish vocabulary for secondary and upper secondary school in the OrdiL-project (Lindberg

and Johansson Kokkinakis, 2007). Another Swedish project modeling different aspects of the lexical knowledge of a language learner in vocabulary assessment is the MOA-project (Lindberg and Johansson Kokkinakis, 2011). In these two projects language pedagogical aspects are emphasized and benefits from focusing on everyday vs. scientific language. Research has so far shown that students with a different language background encounter difficulties with polysemous words, in particular those with subject-specific senses which sometimes also have a more general everyday sense.

Another goal is to incorporate SAWL as a lexical resource in CALL-based platforms cf. the Swedish Lärka <http://spraakbanken.gu.se/larka> which is under development in Språkbanken at the University of Gothenburg.

References

Jorge Baptista, Neuza Costa, Joaquim Guerra, Marcos Zampieri, Maria Cabral and Nuno Mamede. 2010. P-AWL: Academic Word List for Portuguese. Computational Processing of the Portuguese Language, In: Lecture Notes in Computer Science, 6001/2010:120–123.

Laurie Bauer and Paul Nation. 1993. Word families. International Journal of Lexicography, 6:253-279.

Tove Bjørneset. 2001. Introduksjon til ordboksprosjektet NORDLEXIN-N. In: Martin Gellerstam, Kristinn Jóhannesson, Bo Ralph and Lena Rogström (Eds.) Rapport från Konferens om lexikografi i Norden Göteborg 26-29 maj 1999. (Nordiska studier i lexikografi 5.): 44–53. Göteborg.

Tom Cobb and Marlise Horst. 2004. Is there room for an academic word list in French? In: Paul Bogaards and Batia Laufer (Eds.) Vocabulary in a second language: Selection, acquisition, and testing, 15-38. John Benjamins, Amsterdam.

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

Averil Coxhead. 2002. The Academic Word List: A corpus-based word list for academic purposes. In: Bernard Ketteman and Georg Marko (Eds.) Teaching and Language Corpora (TALC) 2000 Conference Proceedings. Rodopi, Atlanta.

John Eldridge. 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.

Nick. C. Ellis, Rita Simpson-Vlach and Carson Maynard, (2008). Formulaic Language in Native and Second Language Speakers: Psycholinguistics, Corpuslinguistiscs and TESOL. TESOL Quarterly 42(3):375–396.

Jaroslava Hlaváčová. 2000. Rarity of words in a language and in a corpus. Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000):1595-1598.

Jaroslava Hlaváčová. 2006. New Approach to Frequency Dictionaries – Czech Example. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006):373-378.

Sylviane Granger and Magali Paquot. 2009. Lexical Verbs in Academic Discourse: A Corpus-driven Study of Learner Use. In: Maggie Charles, Diane Pecorari and Susan Hunston (Eds.) Academic Writing: At the interface of corpus and discourse. Continuum, London/New York.

Ken Hyland and Polly Tse. 2007. Is There an "Academic Vocabulary"? TESOL Quarterly, 41(2):235-253.

Håkan Jansson, Sofie Johansson Kokkinakis, Judy Ribeck and Emma Sköldberg. 2012. A Swedish Academic Word List: Methods and Data. In: Ruth Vatvedt Fjeld and Julie Matilde Torjusen (Eds.) Proceedings of the 15th EURALEX International Congress 7–11 August, 2012. Oslo University, Oslo.

Sofie Johansson Kokkinakis, Emma Sköldberg, Birgit Henriksen, Kari Kinn and Janne Bondi Johannessen. 2012. Developing Academic Word Lists for Swedish, Norwegian and Danish – a joint research project. In: Ruth Vatvedt Fjeld and Julie Matilde Torjusen (Eds.) Proceedings of the 15th EURALEX International Congress 7–11 August, 2012. Oslo University, Oslo.

Adam Kilgarriff, Pavel Rychlý, Pavel Smrz and David Tugwell. 2004. The Sketch Engine. In: Geoffery Williams and Sandra Vessier (Eds.) Proceedings of the Eleventh EURALEX International Congress, EURALEX2004: 123–131, Lorient, France, May 6-10, 2004. Université de Bretagne Sud, Lorient. [online: http://www.euralex.org/elx_proceedings/Euralex2004/011_2004_V1_Adam%20KILGARRIFF,%20Pavel%20RYCHLY,%20Pavel%20SMRZ,%20David%20TUGWELL_The%20%20Sketch%20Engine.pdf].

Lexins svenska lexikon. 2011. 4:th ed. [www] http://lexin.nada.kth.se/lexin/.

Inger Lindberg and Sofie Johansson Kokkinakis. 2007. OrdiL. En korpusbaserad kartläggning av ordförrådet i läromedel för grundskolans senare år. (ROSA, Rapporter om svenska som andraspråk 8.) Institutet för svenska som andraspråk, Göteborgs universitet, Göteborg. [on-line: http://gupea.ub.gu.se/dspace/handle/2077/20503]

Inger Lindberg and Sofie Johansson Kokkinakis. 2011. Identification of lexical cohesive ties in secondary school text books. AILA 2011. The 16th World Congress of Applied Linguistics, Beijing, China.

Krister Lindén and Lauri Carlson. 2010. FinnWordNet – WordNet på finska via översättning. LexicoNordica, 17:119–140.

Nina Martola. 2011. FinnWordNet och kulturbundna ord. LexicoNordica, 18:111–133.

Hilary Nesi. 2002. An English Spoken Academic Word List. In: Anna Braasch and Claus Povlsen (Eds.) Proceedings of the Tenth Euralex International Congress 2002 (vol. 1): 351–357. Center for Sprogteknologi, Copenhagen.. [on-line: http://www.euralex.org/elx_proceedings/Euralex2002/036_2002_V1_Hilary%20Nesi_An%20English%20Spoken%20Academic%20Wordlist.pdf].

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. [on-line: http://www.oecd.org/dataoecd/36/44/38235147.pdf].

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

Petr Savický and Jaroslava Hlaváčová. 2002. Measure of word commonness. Journal of Quantitative Linguistics 9,:215−231.

Mike Scott. 1997. PC analysis of key words – and key key words. System 25/2, p. 233−245.

Rita Simpson-Vlach and Nick C. Ellis (2010). An Academic Formulas List: New Methods in Phraseology Research. Applied Linguistics 31(4):487-512.

Emma Sköldberg and Sofie Johansson Kokkinakis. 2012. A och O om akademiska ord. Om framtagning av en svensk akademisk ordlista. In: Birgit Eaker, Lennart Larsson and Anki Mattisson (Eds.). Nordiska studier i lexikografi 11. Rapport från Konferensen om lexikografi i Norden: 575–585. Lund 24−27 maj 2011.

Michael West. 1953. A general service list of English words: with semantic frequencies and a supplementary word-list for the writing of popular science and technology. Longman, London.

Guoyi Xue and Paul Nation. 1984: A university word list. Language Learning and Communication 3(2):215–229.