Introducing the Swedish Kelly-list, a new lexical e-resource for Swedish

Elena Volodina, Sofie Johansson Kokkinakis

University of Gothenburg, Sweden
Språkbanken, Institutionen för svenska språket, Göteborgs universitet, Box 200, 405 30
Göteborg
elena.volodina@svenska.gu.se, sofie.johansson.kokkinakis@svenska.gu.se

Elena Volodina, Sofie Johansson Kokkinakis

Frequency lists and/or lexicons contain information about the words and their statistics. They tend to find their "readers" among language learners, language teachers, linguists and lexicographers. Making them available in electronic format helps to expand the target group to cover language engineers, computer programmers and other specialists working in such areas as information retrieval, spam filtering, text readability analysis, test generation etc.
This article describes a new freely available electronic frequency list of modern Swedish which was created in the EU project KELLY. We provide a short description of the KELLY project; examine the methodological approach and mention some details on the compiling of the corpus from which the list has been derived. Further, we discuss the type of information the list contains; describe the steps for list generation; provide information on the coverage and some other statistics over the items in the list. Finally, some practical information on the license for the Swedish Kelly-list distribution is given; potential application areas are suggested; and future plans for its expansion are mentioned. We hope that with some publicity we can help this list find its users.

Keywords: frequency lists, corpus-based approach, lexical e-resource

1. Background

1.1 On KELLY project

The Swedish Kelly-list was produced as a result of the project KELLY (KEywords for Language Learning for Young and adults alike), funded by the EUs Lifelong Learning Programme, KA2 Languages subprogramme. It was granted to Stockholm University as project co-ordinator with eight academic and enterprise partners (http://su.avedas.com/converis/contract/321) for two years starting in November 2009.
The aim of the project was to create a language learning tool to be used in nine different languages; Arabic, Chinese, English, Greek, Italian, Norwegian, Polish, Russian, and Swedish. The tool consists of around 9000 keywords of each language described by frequency and language proficiency level. The model used in creating frequency lexicons is influenced by the Common European Framework of Reference (CEFR; Council of Europe, 2001).
Another aim of the KELLY project was to fill a gap and a need for language learning resources for learners of the nine partner languages. There are a lot of programs and courses nowadays that are evaluated in terms of levels described in CEFR. The curriculum for CEFR-oriented courses, however, needs further development (Little, 2011). Attempts have been made to identify how many hours each CEFR level can demand in teacher-driven education or approximately how many words per level and sublevel (Deutsche Welle). However, up to date there has never been any description of which vocabulary learners of each CEFR level should master, or how many words each level should consist of, which provides another reason for using the Swedish Kelly-list.

1.2 Existing frequency-based lists for Swedish

Frequency dictionaries contain most frequent words in one's language together with information on how often the words are used in the source material. This information often provides guidelines as to which words are most important for language learners, which headwords should be included into dictionaries in the first place, provide material for linguistic and comparative analysis.
Different frequency lexicons may contain different information. One of the frequency lexicons over Swedish, "Nusvensk frekvensordbok baserad på tidningstext" (Eng. "Frequency Dictionary of Contemporary Swedish Based on Newspaper Texts") (Allén, 1970) comes in four volumes, the first one based on graphic words; the second one on lemmas; the third on collocations and the last one on morphemes. "A Frequency Dictionary of Spanish. Core Vocabulary for Learners" by Mark Davies (2011) contains apart from the headword and its frequency translation into English, example sentences, and an indication of major register variation.
Swedish tradition of frequency lexicons is presented by a few dictionaries, e.g. Sture Allén's "Tiotusen i top" (1972) and "Nusvensk frekvensordbok baserad på tidningstext" in four volumes (1970); Kent Larsson, Carin Anderson and Valentina Rosén's "Frekvensordbok över svenska elevtexter" (Eng. "Frequency Dictionary of Learner Texts") (1985); Jens Allwood's "Talspråksfrekvenser" (Eng. "Colloqual Language Frequencies") (Allwood, 1999). All of the above-mentioned sources are available in paper format and have copyright restrictions.
Recently a frequency list over the Swedish base vocabulary was compiled by Eva Forsbom (2006). This resource, called Base Vocabulary Pool, is openly available for use and presents a frequency-based word list of 8,215 lemmas constituting central Swedish vocabulary derived from the

SUC (Stockholm Umeå Corpus), SUC being a balanced corpus of written Swedish of 1990-s.
The above-mentioned lists present a very rich material for comparison, evaluation and potential enrichment of the Swedish Kelly-list.
The main distinction that the Swedish Kelly list possesses compared to the above-mentioned frequency lists is that it reflects modern language (as of 2010) and has been collected on an extensive text data from web which (1) ensures a mixture of genres, i.e. texts not biased towards any specific domain and (2) includes several language modes, both written and spoken-like (blogs, forums, chats).

2. The Swedish Kelly-list

2.1 Information on the list

The Swedish Kelly-list introduced in this paper is a freely available frequency-based vocabulary list that comprises general-purpose language; it is generated from a large web-acquired corpus (SweWAC) of 114 mln. words from the 2010's. Further, it is adapted to the needs of language learners and contains 8,425 most frequent lemmas that cover 80% of SweWAC (Johansson Kokkinakis, S. & Volodina, E., 2011).
The headwords on the Swedish Kelly-list contain the following information, see also Table 1:
(1) id/running number (i.e. relative placement in the frequency band);
(2) raw frequency (RF);
(3) relative frequency , i.e. "word-per-million" (WPM);
(4) CEFR level (A1, A2, B1, B2, C1, C2);
(5) source of lemma[1] (indication whether the headword[2] comes from SweWAC, from translation list (T2) or has been manually added);
(6) grammar information, in our case limited to articles and infinitive markers, added so that students could differentiate between neuter and non-neuter nouns as well as between modal and content verbs;
(7) lemma, sometimes provided together with its spelling/stylistically marked variant;
(8) part of speech[3];
(9) comments/examples for some of the headwords, mostly coming from the candidate lists for inclusion (see subsection 2.5) from the languages where multiple word units have been included into the monolingual lists.

ID	RF	WPM	CEFR lev	Source	Gram	Item	POS	Example
88	2,624,032	23,017.26	A1	Swe-WaC	att	vara (var-dagl. va)	verb	e.g. var så god!

Table 1. Example of items in the Swedish Kelly-list

The entry (row) should be read in the following way: the verb *att vara* (Eng. *to be, to last*) has a colloquial variant *va*; it can be used in a phrase *var så god!* (Eng. *here you go!*); it has the rank "88" in the list and thus belongs to the language's top 100 words. It has been used 2,624,032 times in SweWAC (RF) which gives 23,017.26 wpm value. The item belongs to the most important vocabulary for language learners and should be learned at A1 CEFR level.

2.2 Corpus compiling

The Swedish Kelly-list was created in a five-step process as described in Johansson Kokkinakis & Volodina (2011). The steps are outlined briefly here in subsections 2.2-2.5.

The first step consisted in compiling a new corpus. In order to make the frequency lists for the nine partner languages comparable, they had to be derived from modern web corpora of approximately equal size. The minimal size restriction was set to 100 mln. words, since large corpora provide reliable statistics over the word usage. Unfortunately, there was no corpus of Swedish of that size before the project start. To settle that problem, a web-corpus SweWAC (Swedish Web Acquired Corpus) has been collected by the KELLY partner "Lexical Computing Ltd" using Corpus Factory tool (Kilgarriff, Reddy, Pomikálek, 2010).
Compiling a web-based corpus for Swedish was a process consisting of several steps:
(1) Collect "seed word" list, approximately 500 mid-frequency words whose frequency range is between 1000 and 6000. This was done using texts on Wikipedia – first a "Wiki-corpus" was compiled as a primary corpus for seed-word selection, word form frequency was calculated (as opposed to base forms/lemmas), and then 500 mid-frequency word forms were selected for further web-search. Length restriction was set on the seed words: they should be at least 5 characters long to sort out coinciding word forms in other languages (e.g. Swedish versus English *fast*). Words containing digits or other non-characteristic for the language characters were discarded.
(2) Repeatedly select three random seed words to create a query, and send the query to a search engine.
(3) Retrieve hit pages and clean the text, e.g. remove navigation bars, ads, duplicates. The web-corpus finally consisted of 114 million words.

The raw texts have been handed over to the Swedish Language Bank where they were annotated for parts of

[1] Lemma in this article is understood as a base form plus its part-of-speech tag.
[2] Headwords in the Swedish Kelly list encompass the following items: single-word base forms; a number of multiple word items identified automatically during the POS-tagging phase as well as reflexive verbs that have been assigned the reflexive pronoun during the manual proofreading stage; some abbreviations.
[3] *Part of speech*, *POS* and *word class* are used interchangeably as synonyms in this article

speech and lemmas using tools available through the Swedish Language Bank (Kokkinakis and Johansson Kokkinakis, 1997).

Among the advantages of web-collected corpora we can name the following:
- Since its construction is a highly automated process, short collection time at low costs is ensured.
- Texts collected from the web tend to contain more spoken-like interactional language since there are a lot of forums and blogs; therefore, compared to classical corpora, they have a benefit of complementing strictly written mode of language with everyday colloquial language.
- Texts represent modern language.
- Since the text data is extensive, a reliable information over the word usage can be derived.

Among the disadvantages or rather limitations of a web corpus we can name the following:
- First of all the absence of control over the kinds of texts that constitute the corpus. Such corpora are therefore unpredictable as to their structure and contents, presenting an unclear mixture of domains and most probably devoid of balance between domains and genres. However, after the analysis of the Kelly-list and SweWAC we came to a conclusion that there are three predominant text types in SweWAC: political texts, web- and computer-related texts, as well as online communication (blogs, chats, forums).
- As our experience of SweWAC has shown, besides texts in Swedish there is a minor percentage of texts written in other languages, among them Norwegian, Danish and English. Presumably the reason for that is presence of ambiguous seed words, for example international proper names, e.g. *Albert*, *Alexander*, *Berlin*, *Chris*, *Chicago*, *Daniel*; non-Swedish spelling of words, e.g. *America* (as opposed to the Swedish *Amerika*), British (as opposed to *brittisk*), *company* (Swedish *företag*), *college*, *corporation* etc. A number of seed words coincided in form with English words, even though their length was longer than or equal to five characters, e.g. *album*, *attack*, *civil*. One way out of this may be POS-tagging of the wiki-corpus and filtering seed words of unwanted word classes (e.g. proper names and foreign words) prior to sending queries to the search engines. Another even better alternative is to have a language team prepare a list of seed words (or even better several lists for different genres/domains) and thus ensure the more or less balanced and predictable structure of the corpus.

However, these limitations have proven to be minor problems. The method of working on the KELLY-lists was formed in such a way that most of the problems mentioned above were corrected during the validation phase through word list comparisons between languages. This and some other selection strategies are described later in this article.

SweWAC is at present available in its original form via commercial concordance tool SketchEngine <http://www.sketchengine.co.uk/> as well as in the form of a "citation" corpus, in which sentences are mixed in random order so that the full texts cannot be retrieved via the freely available concordance tool Korp, developed and distributed by the Swedish Language Bank <http://spraakbanken.gu.se/korp/>.

2.3 List generation

The second step involved applying information on statistics in the SketchEngine (Kilgarriff et al., 2004), including generation of the first frequency list. There are three frequency measures that have been used in the Swedish Kelly-list: raw frequency (RF), relative frequency (word per million or WPM) and average reduced frequency (ARF). Raw frequency gives an absolute count of the words in the corpus. WPM is the relative count where raw frequency is divided by the total number of running words (tokens) in the corpus and then multiplied by one million. WPM is a measure which makes word frequencies from different sources/corpora comparable. ARF takes into account dispersion of the words in different subcorpora and throughout the whole corpus. If the word/lem-pos is used in only one of the subcorpora, or if the distance between the word occurrences in the whole corpus is not regular, it is not considered to be representative of the basic vocabulary, and its rank is reduced according to the formula explained in Savický and Hlavácová (2002). The measure is used to ensure that only domain-independent general-purpose vocabulary is selected, i.e. words that are frequent in a few texts of a certain domain (e.g. law or medicine) but otherwise not regularly used in all types of texts are disqualified from the general vocabulary status.

The first version of the monolingual frequency list (M1) was selected by ARF, and afterwards ordered by WPM.

The main guidelines in selecting headwords for the KELLY-lists were developed collectively by the partners in the form of a document "Proposal for inclusion of word types in Kelly". According to those guidelines each language team should include lem-pos with normalized spelling, avoid "language-family" principle, i.e. include derivational forms as legitimate independent items; avoid including idioms or other phraseological units; avoid proper names with a few exceptions. Homonymy, polysemy, multiword expressions (mwe) and abbreviations were left for each language team to decide upon.

The following word classes were suggested for inclusion: noun, verb, adjective, adverb, pronoun, determiner, conjunction (and subjunction), exclamation and some numerals, namely: 1-20, 30, 40, 50, 60, 70, 80, 90, 100, 1000, 1000000, 1st, 2nd, 3rd(but not 4th, 5th, ...), half, quarter, third.

The following word classes were suggested for exclusion: participle, proper nouns, foreign words, punctuation. The main principle in selecting candidate parts of speech was relevance for language learners and feasibility for translation mappings between languages.

The M1 list was filtered for noise, defined by us as:
- entries containing digits, and non-letter characters, such as ><=;*'/ etc.);

- some word classes: punctuation, proper names, numerals (except the ones mentioned above), participles, foreign words.
This step produced a lemma-list of 54 000 items.

2.4 From M1 to M2. Proofreading

The third step was "lexicographic" in character. The top 9 000 candidates for headwords were selected. The list needed correction of various lemmatizing and tagging errors; for example the noun *fånge* (Eng. *prisoner*) was erroneously lemmatized as a non-existent noun *fångare* from its definite plural form *fångarna*. We merged words with different style and spelling variants into the same headword entry. In some cases we consulted SAOL online (<http://www.svenskaakademien.se/svenska_spraket/svenska_akademiens_ordlista/saol_pa_natet/ordlista>) before we made decisions on, for example, which variant should be made headword and which one provided in brackets as an alternative variant.
Once the first manual work was completed, the list was automatically proofread by matching headwords against SALDO, an electronic lexicon of Swedish (Borin, Forsberg, 2009), and against SMDB, the Swedish Morphological Database (Berg & Cederholm, 2001; Johansson Kokkinakis, 2001). Finally the list was manually proofread, consistency for item presentation checked, article and infinitive markers assigned. This step left us with 8,484 headwords including 83 manually added items relevant for learners of Swedish.

2.5 Validation through translation

After the top 6 000 headwords were translated into eight partner languages by a translation agency, the "validation-through-translation" stage started.
The KELLY database (Kelly DB, http://kelly.sketchengine.co.uk/) was created by the LCL-partner for reasons of comparison of translation lists versus original monolingual lists which resulted, among other things, in lists of items recommended for deletion and inclusion. To make an example, all translation lists into Swedish (8 of them) were compared with the Swedish monolingual list M2, and this generated a list of items that appeared in the M2 list but never in the translation lists (deletion candidates); as well as a list of items present in translation lists but not present in the Swedish M2 list (inclusion candidates). On the basis of those lists the Swedish monolingual list was reanalyzed and finalized embedding linguistic evidence from the translation lists paying special attention to learner relevant domain-specific words (Volodina and Johansson Kokkinakis, 2012). The Swedish monolingual list then consisted of 8,425 entries ordered after the WPM frequency.
As a side effect, the Kelly DB facilitated generation of a number of other lists, namely universal vocabulary (i.e. items present in all the 9 languages); common vocabulary lists (i.e. vocabulary shared by 8, 7, 6, etc. Languages); words specific to each individual language pair; as well as unique vocabulary (i.e. vocabulary specific to each individual language). The first explorations of Kelly DB is presented in Kilgariff et.al. (submitted); some description is provided in Johansson Kokkinakis et.al. (2011) and Volodina et.al. (2012).

3. Coverage

3.1 General on vocabulary distribution in the Swedish Kelly-list

The 8,425 headwords on the Swedish Kelly-list have been equally assigned to CEFR levels according to their frequency range, approx. 1,404 headwords per level.
With respect to their sources, the headwords are distributed in the following way:
- 85 have been added manually. They constitute 1% of the list, all belonging to CEFR A1 and cover 0,44% of SweWAC.
- 2,564 headwords come from T2 (translation lists). They constitute 30,4 % of the Kelly-list and cover 1,7% of SweWAC texts. Approximately 2,500 of those items appear in the last two proficiency levels C1 and C2, as shown in table 2.
- 5,776 headwords come from SweWAC. They constitute 68,5 % of the Kelly-list and cover 77,98% of the total SweWAC texts. They appear evenly in the first four CEFR levels, and disappear at all from the last CEFR level C2, as shown in table 2.

CEFR level	**Nr of T2 words**	**SweWAC coverage, %**	**Nr of SweWAC items**	**Swe-WAC cover-age, %**
1 (A1)	14	0.7	1,305	68.9
2 (A2)	27	0.0909	1,377	5.3198
3 (B1)	53	0.0882	1,351	2.26
4 (B2)	69	0.12	1,335	1.16
5 (C1)	996	0.495	408	0.2686
6 (C2)	1405	0.2476	0	0
Total	**2,564**	**1.6739**	**5,776**	**77.98**

Table 2. SweWAC coverage by T2 and SweWAC items.

Another interesting piece of statistics is the distribution of parts-of-speech in the Swedish Kelly-list and their coverage. Table 3 presents some parts of speech (the coverage number is given in percent of the total number of tokens in SweWAC):

POS	Total count (% of Kelly-list)	Coverage, SweWAC
Adjective	1,354 (16.07%)	6.43%
Adverb	569 (6.75%)	7.6%
Determiner	10 (0.12%)	3.6%
Noun	4,607 (54.68%)	14.51%
Preposition	108 (1.28%)	11.14%
Pronoun	61 (0.72%)	11.4%
Verb	1,538 (18.26%)	16.9%

Table 3. Kelly POS distribution in SweWAC

A curious observation can be made that 61 pronouns cover 11.4% of SweWAC; 108 prepositions make up 11.14%; whereas 4,607 nouns cover only 14.51% compared to 1,538 verbs that cover 16.9%. Knowledge of verbs, pronouns and prepositions appears more "beneficial" than knowledge of nouns in terms of text coverage, or so it would seem from statistics. That ironic conclusion compromises in general statistically-based conclusions unless a sound analysis of the aspects that the statistical calculations are based upon is made.

3.2 Corpora coverage by Kelly-items

Words from the Swedish Kelly-list cover 80% of the lexical items of SweWAC. Punctuation marks constitute next 10% of the corpus. Table 4 presents SweWAC coverage of lexical items per CEFR level by Kelly items:

CEFR level	A1	A2	B1	B2	C1	C2	**Total, %**
Cover-age, %	70	5.5	2.3	1.3	0.8	0.3	**80.1**

Table 4. SweWAC coverage by Kelly items per CEFR level

We have performed coverage tests on three corpora: the core corpus SweWAC, and two control corpora - Parole and SUC.

Both Parole and SUC are well-annotated general-purpose corpora of written Swedish. Texts in Parole date from 1976-1997 and comprise newspaper texts and imaginative prose. SUC dates from 1990-s, and is a balanced corpus of written language coming from 9 genres. SUC has been semi-automatically tagged, all texts have been afterwards manually proofread.

The coverage calculations have shown that words from the Swedish Kelly-list cover 80% of the total of SweWAC, punctuation, infinitive markers and proper names stand for the next 16%. It looks very encouraging. However, coverage calculations over the two other control corpora have shown a less encouraging result: Kelly words cover only 62.75% of Parole corpus and 68.87% of SUC as can be seen in table 5.

Parameter	**Swe-WAC**	**Parole**	**SUC**
Punctuation, coverage %	10.7%	12.7%	11.5%
Infinitive marker, coverage %	1.26%	1.01%	1.1%
Proper names, coverage %	4.87%	8.67%	3.6%
Kelly-words, coverage %	79.65%	62.75%	68.87%
Total coverage	96.5%	85.14%	85.07%

Table 5. SweWAC, Parole and SUC coverage.

A number of Kelly-items got zero-matches in the control corpora: 653 items didn't appear at all in SUC and 224 had no match in Parole. Such a drop in coverage percentage as well as zero-matches are surprising at first glance. A short check, however, has revealed the reasons for this drop in values that can be summarized as follows: (1) differences in tagging and lemmatization; and (2) difference in text genres constituting the three corpora.

(1). Lemmatization and pos-tagging of the two control corpora differ from the SweWAC-based Kelly-list. The headwords in the Kelly-list have undergone manually introduced changes. As a result a number of items were corrected for POS tags or lemma, for example *själv* (Eng *self*) changed its tag from *adjective* to *pronoun*. Tagging differences can also be seen in POS-mismatches in such highly frequent words as *ett*, *det*, *sin*, *annan*, etc. that are tagged as *pronouns* in the Kelly-list as opposed to *determiner* in SUC.

Furthermore, a number of headwords in the Kelly-list have been modified to make them more user-friendly for L2 learners. For example, reflexive verbs (e.g. *te_sig*) where the reflexive pronoun *sig* has been added during the proofreading stage.

(2). The second difference lies in the type of texts used in different corpora. Since SweWAC is a web corpus of more modern language than SUC or Parole, it shows lexical development of the recent decade. For instance, the following groups of vocabulary are present in SweWAC but absent in SUC and/or Parole (see more in Johansson Kokkinakis et.al. 2011 and Volodina et.al. 2012):

(1) vocabulary reflecting recent "hot" political events, e.g. *piratparti*, *svininfluensa*, *alliansregering* (Eng. *pirate party*, *swine flu*, *alliance government*);

(2) vocabulary related to web- and computer-related domains, e.g. *blogga, bloggare, webbläsare,* (Eng. *to blog, a blogger, web browser*);
(3) vocabulary colloquial in nature characteristic of online conversation present in SweWAC (blogs, chats, forums), e.g. *toppen, jävla, tryne* (Eng. *great, damn, snout*);
(4) down-to-earth learner-specific domain vocabulary, such as *socka, huva, sparv, aprikos, brorsdotter* (Eng. *sock, hood, sparrow, apricot, niece*);
(5) and, finally widely spread loaned words such as *shopping, klick, mejl, kidnappning, designer, server*.

This type of check has confirmed our hypothesis about the text genres that are typical of SweWAC, namely newspaper texts, web- and computer related texts as well as blogs and forums.
To sum it up, we can claim that, had it not been for lemmatization and POS-tagging mismatches, the coverage numbers would have been increased for both Parole and SUC. Moreover, the vocabulary absent in SUC and Parole as shown in (2) above is both modern and relevant vocabulary for L2 learners.
Thus, assuming that the learner who knows words from the Swedish Kelly-list would have no difficulty coping with punctuation and infinitive markers, his/her vocabulary competence will allow understanding of approximately 80-90% of the modern Swedish texts.

4 Some final comments

The way the Kelly-list is compiled, it is a reliable resource for suggesting lexical syllabus for CEFR-based courses in Swedish as well as for use in evaluating learner appropriate texts for different CEFR levels, for compiling course books, creating vocabulary exercises and tests, compiling dictionaries, and for a number of other language learning purposes and NLP applications. The list can be used by language learners and teachers, test creators, lexicographers, comparative linguists, corpus linguists, computational linguists, and many other user groups.
Apart from representing the most frequent core vocabulary of modern Swedish derived from a large web-acquired corpus, the Swedish Kelly-list is based on objective selection, i.e. human judgment was avoided in favor of objective decisions. The word selection has been strictly frequency-based with a few cases of pedagogically grounded modifications, additions and deletions. Even the latter ones followed straightforward principles so that the experiment with the Swedish Kelly-list can be reproduced.
The Swedish Kelly-list is a freely available electronic resource and is distributed under the license agreement CC-BY-SA 3.0, LGPL 3.0. It can be downloaded from <http://spraakbanken.gu.se/eng/kelly>. You are encouraged to make a reference to this or any other article describing this list if you use the Swedish Kelly-list.
We have plans for further expansion and exploitation of the Swedish Kelly-list, among other things creation of a dynamic lexical database with a possibility for selecting lists of domain words, for adding corpus examples and translation equivalents. Linking this resource to other lexicons available through the Swedish Language Bank we can get morphological analysis of the headword items, their monolingual definitions, and a number of other interesting options. Test item generation as well as lexical analysis of text complexity can also be named among future plans of exploitation of this list.

5 References

Allén, S. (1970). *Nusvensk frekvensordbok på tidningstext.*. Almqvist & Wiksell, Sweden.

Allén, S. (1972). *Tiotusen i top*. Almqvist & Wiksell, Sweden.

Allwood, J. (1999). Talspråksfrekvenser. In *Gothenburg Papers in Theoretical Linguistics S20*. Revised and updated version (1999). Gothenburg Papers in Theoretical Linguistics, S21, Göteborg University. Dept of Linguistics. pp. 1-416.

Berg, S., Cederholm, Y. (2001). Att hålla på formerna. Om framväxten av Svensk morfologisk databas [On the creation of the Swedish Morphological Database]. In: *Gäller stam, suffix och ord*. Publication in honor of Martin Gellerstam, October 15, 2001. Meijerbergs arkiv för svensk ordforskning 29. pp. 58-69. ISBN: 91-631-1542-5.

Borin, L., Forsberg, M. (2009). All in the family: A comparison of SALDO and WordNet. *Proceedings of the Nodalida 2009 Workshop on WordNets and other Lexical Semantic Resources - between Lexical Semantics, Lexicography, Terminology and Formal Ontologies*. Odense.

Council of Europe (2001). *The Common European Framework of Reference for Languages*. Cambridge University Press

Davies, M. (2011). A *Frequency Dictionary of Spanish. Core Vocabulary for Learners*. Taylor & Francis, London, Great Britain.

Deutche Welle. *Deutschkurs*. Retrieved from http://deutschkurse.dw-world.de/dw_static_content/langerklaerung_en.html Access date: June 25th 2011

Forsbom, E. (2006). *Deriving a Base Vocabulary Pool from the Stockholm Umeå Corpus*. Retrieved from http://stp.lingfil.uu.se/~evafo/resources/basevocpool/ Access date: January 6th 2011

Johansson Kokkinakis, S. (2001). Disambiguering av homografa ord i Språkbanken med hjälp av Svensk morfologisk databas. [Disambiguating homographs using the Swedish Morphological Database]. In: *Gäller stam, suffix och ord. Publication in honor of Martin Gellerstam*, October 15, 2001. Meijerbergs arkiv för svensk ordforskning 29. pp. 177-188. ISBN: 91-631-1542-5.

Johansson Kokkinakis, S. and Volodina, E. (2011). Corpus-based approaches for the creation of a frequency

based vocabulary list in the EU project KELLY – issues on reliability, validity and coverage. *eLex 2011*, Slovenia.

Kilgarriff A., Charalabopoulou F., Gavrilidou M., Bondi Johannessen J., Khalil S., Johansson Kokkinakis S., Lew R., Sharoff S., Vadlapudi R, Volodina E. (submitted, LREJ 2012). Corpus-Based Vocabulary lists for Language Learners for Nine Languages. *LREJ special issue*.

Kilgariff A., Reddy S., Pomikálek J., PVS Avinesh (2010). A Corpus Factory for Many Languages. *Proceedings of LREC 2010*

Kilgariff A., Rychly P., Smrz P., Tugwell D. (2004). The Sketch Engine. *Proc EURALEX 2004*, Lorient, France.

Kokkinakis, D. & Johansson Kokkinakis S. (1997). *A Robust and Modularized Lemmatizer/Tagger for Swedish Based on Large Lexical Resources*, Inst. F. svenska språket, Göteborgs universitet.

Larsson K., Anderson C., Rosén V. (1985). *Frekvensordbok över svenska elevtexter*. FUMS and UDCL, Uppsala, Sweden

Little D. (2011). The common European Framework of Reference for Languages: A research Agenda. *Language Teaching*, vol. 44.3, p.381-393. Cambridge University Press.

Savický P. and Hlavácová J. (2002). Measures of word commonness. *Journal of Quantitative Linguistics*, *9*: 215-231.

Volodina E. and Johansson Kokkinakis S. (2012). Swedish Kelly: Technical Report. GU-ISS 01-12. The Swedish Language Bank, Gothenburg University.