SweVoc - A Swedish vocabulary resource for CALL

Katarina Heimann Mühlenbock and **Sofie Johansson Kokkinakis**

Dept of Swedish, University of Gothenburg, Gothenburg, Sweden
katarina.heimann.muhlenbock@gu.se, sofie@svenska.gu.se

Abstract

The core in language teaching and learning is vocabulary, and access to a delimited set of words for basic communication is central for most CALL applications. Vocabulary characteristics also play a fundamental role for matching texts to specific readers. For English, the task of grading texts into different levels of difficulty has long been facilitated by the existence of word lists serving as guides for vocabulary selection. For Swedish, the situation is with a few exceptions less fortunate, in that no base vocabulary organized according to aspects of usage has existed. The Swedish base vocabulary – SweVoc – is an attempt to remediate this. It is a comprehensive resource, aimed at differentiating vocabulary items into categories of usage and frequency. As we are of the opinion that no corpus of written text can do fully justice of general language use, we have utilized materials from a second language as reference for delimiting the category of core words. Another belief is that the task of defining a base vocabulary can not be fully automatic, and that a considerable amount of manual, traditional lexicographic work has to be invested. Hence, the present approach is not an innovative, but a methodological approach to word list generation for a specific purpose, much like LSP. We anticipate SweVoc to be integrated in CALL applications for vocabulary assessment, language teaching and students' practice.

1 Background

Vocabulary knowledge plays a central role in a person's ability to communicate, as well as reading and understanding written text. It is therefore a central issue in many readability assessment approaches. Prominent researchers within readability and language assessment, such as (Thorndike, 1921; Vogel and Washburne, 1928; Patty and Painter, 1931; Thorndike and Lorge, 1944; Dale and Chall, 1948; Spache, 1953), and more recently (Nation, 1990; Nation, 2001), all included specific word lists as a criterion to measure text difficulty for English. In quantitative associative studies of readability, some scheme for measuring the vocabulary difficulty is set up, compared to a predefined criterion, and expressed by a coefficient of correlation. In this way, the word lists may be constructed in order to mirror vocabulary difficulty corresponding to school grade levels. Thorndike's (1921) word list of 10,000 words, later on revised into a list of 30,000 words (Thorndike and Lorge, 1944) and Spache's revised word list (Spache, 1974) of 1,040 entries, were mainly constructed by judgment and common sense. West published in 1953 the General Service List – a list of 2,000 words selected to represent the most frequent words in an English corpus.

Vocabulary is also an important issue when producing language-supportive aids for persons with deficient communication capability. Insufficient vocabulary knowledge implies a decrease in expressive power of an utterance or written text, and the receptive language skills are also heavily dependent upon the individual vocabulary range. In order to obtain

Katarina Heimann Mühlenbock and Sofie Johansson Kokkinakis 2012. SweVoc – A Swedish vocabulary resource for CALL. *Proceedings of the SLTC 2012 workshop on NLP for CALL.* Linköping Electronic Conference Proceedings 80: 28–34.

maximum benefit from language supportive tools, the resources provided as word lists ought to be chosen with care in order to conform to individual and situational needs. Also in generating LSP (language for specific purposes) and particular domain vocabulary lists, a list of general base vocabulary is needed in order to exclude the most common and general words.

In the following we are making a distinction between *base vocabulary* and *core vocabulary*. A language teaching situation might involve a more extensive base vocabulary, while assistive technology applications such as symbol boards for communication would benefit from a restricted core vocabulary, expandible with complementary vocabulary items from different domains. The present approach is an attempt to combine both models, i.e. it is a Swedish core vocabulary word list, supplied with words belonging to a broader base vocabulary.

Defining a core vocabulary is a task associated with several methodological challenges. Lee (2001) has enumerated some of them. First of all, the concept of *core vocabulary* has to be settled. Several working definitions exist, out of which the most contested point seems to be whether the list is based on, and intended for, applications within written or spoken language, or both. If one decides to adopt the view that a core vocabulary is by definition that which is central to the language as a whole, it rules out for instance approaches based on frequency countings of words in written language. Furthermore, it should be untarnished from any stains of genre, style, register or lect association.

In addition to the theoretically founded issues, also problems of more practical nature arise. Although a major part of verbal communication is said to take place with the use of 1,500 - 2,000 words (West, 1953), this figure must be considered in the light of language-specific properties, of the type of communication, and above all, as a function of the *word* concept. Counting lexemes, lemmas, baseform orthographic words or multiwords render different figures. For English, the notion of *word family* plays a central role when defining word list for educational purposes. Lee (2001), citing Schmitt (2000) maintained that

> people in the field seem to agree that the "word family" is the most meaningful unit to work with and pedagogically most useful.

The word family concept was put forth by Bauer and Nation (1993), from a reader's perspective defined to comprise

> a base word and all its derived and inflected forms that can be understood by a learner without having to learn each form separately.

If all the lemmas belonging to a specific word family are considered as one member of the word list, Hirsh and Nation (1992) found that a vocabulary size of at least 5,000 entries were needed in order to read unsimplified fiction texts. The same study also showed that graded readers beginning at a level of 2,600 word families would be of great benefit in language teaching.

An attempt to construct a levelled base vocabulary for another language than English was made by De Mauro (1980) when he published a list of 7,400 Italian words, categorized into three different groups according to use. The only attempt in this direction for Swedish was made by Forsbom (2006), who derived a base vocabulary pool from a corpus of 1 million words – the Stockholm-Umeå Corpus (SUC) (Källgren, 1992). This was achieved by ranking base word forms according to adjusted frequency over the entire corpus, and then adopting a subsequent filtering technique that sorted out entries which did not occur in more than three out of nine genres in the corpus. The result was a Swedish base vocabulary pool (henceforward referred to as SBVP), with a total amount of $\approx$ 8,200 word base forms, mirroring the use of written Swedish in the early nineties.

SBVP alone neither be considered to reflect modern language use, nor to be enough informative to independently serve as a source of words pertaining to a restricted core vocabulary, since it is based solely on written language. As already mentioned, the base word forms in SBVP are ranked according to adjusted frequency (AF: see equation 1), i.e. relative frequency weighted with dispersion over the 9 categories (genres) in SUC. It implies that the vocabulary are those words that are not genre dependent, given the subdivisions of a small-size text corpus. Furthermore, it lacks information at the lexeme

level, which reduces its feasability for purposes demanding a semantic disambiguation between words. A base form word like the Swedish noun *gång* has for instance four lexeme representations, belonging to different base vocabulary categories. The first refers to 'time' and is considered to be a core vocabulary item, while the sense 'path' is not. The second Issues regarding a distinction between lemma and lexeme concepts are discussed in Gardner (2007). Another flaw in SBVP is the absence of internal levelling, which would be required in order to serve as a list of core vocabulary words. In the present approach, it was hence enriched with labels indicating levels of general use from three additional sources; (1) a translated base vocabulary, (2) a list of words from modern vocabulary, and (3) a dictionary of words denoting domestic life activities and participation in community activities. The final product is SweVoc, a base vocabulary word list, consisting of ≈ 8,500 words, mainly lemma forms, divided into five different categories.

$$AF = (\sum_{i=1}^{n} \sqrt{d_i x_i})^2$$

where
AF = adjusted frequency
d_i = relative size of category i
x_i = frequency in category i
n = number of categories

(1)

2 Material

SweVoc is a comprehensive resource, based on lists of lexical items and texts from four different sources:

1. The backbone was the monolingual Swedish base vocabulary pool (SBVP) (Forsbom, 2006), derived from the SUC corpus (Källgren, 1992), containing 8,213 base form entries. Personal nouns, numbers and punctuation marks were omitted, which reduced the number of entries to ≈ 7,400.

2. The second major resource is a translation of the earlier mentioned work by (De Mauro, 1980), *Guida all'uso delle parole* henceforward referred to as *GUP*. It consists of a vocabulary of 7,400 words, mainly lemma forms, divided into three categories:

   - 2,100 basic words, regarded as fundamental for communication, representing a core vocabulary (C)
   - 2,400 words used in every-day communication (D)
   - 2,900 words highly frequent in written text (H)

3. The *Kelly* modern vocabulary list (Johansson Kokkinakis and Volodina, 2011) was used in order to ensure that frequent words used in modern settings were included. The Swedish version of Kelly is derived from a large modern corpus of web texts, and a subset of ≈ 500 words translated between Swedish and Italian was employed.

4. The *ICF* (Socialstyrelsen, 2003) is a classification of health and health-related domains, ranging from body structure to individual and societal issues. It was used as a reference word list in order to ensure coverage of words related to every-day matters.

3 Preprocessing

The Italian list of basic words was translated into Swedish by a second-language-speaker of Italian. Localisms and archaisms in the source language were ignored. The reason for using a foreign resource was two-fold; First, a base vocabulary should be selected in order to cover both universal concepts and essential phenomena and situations in the main local environment. Secondly, the manual translation task revealed ambiguities due to different usage of words and word senses among two syntactically and lexically distant languages, which contributed to a more fine-grained levelling of words into different subcategories.

The Kelly modern word list was a result of the EC-financed project Kelly <http://kellyproject.eu>. The aim of the project was the generation of monolingual word lists of nine languages, Arabic, Chinese, English, Greek, Italian, Norwegian, Polish, Russian and Swedish. The lists were generated from

many sources including web corpora in order to reflect a modern vocabulary. The lists were then all translated into the eight other languages, generating 72 language pairs. The Italian-Swedish is one of them. The lists were then finally merged to 36 lists. These lists are used in the Keewords language learning tool <http://Keewords.com>.

Several structural differences between the two main sources – SBVP and GUP – caused problems already at the preprocessing stage of SweVoc. As is shown in table 1, the tag set used in SBVP is in PAROLE-format with morphosyntactic information, while GUP was based simply on part-of-speech. In addition to automatic conversion into SUC-format part-of-speech labelling, a considerable amount of manual work was required to make the lists comparable. However, as mentioned already in the introduction, we are of the firm view that no wordlist aimed at specifying a base vocabulary can be produced without a considerable degree of human intervention. We hence regard the present approach to be a pragmatic and feasible way to perform a restricted task.

4 Word list compilation

Entries in SBVP were checked against GUP in order to find candidates for inclusion into SweVoc. As already mentioned, the lists were comparable in size (≈ 7,400 words), but differed largely as regards to compilation methods and contents. As was expected, many words in the each of the two lists corresponded to multiple entries in the other. Multiword expressions and structural differences between the languages also required particular consideration.

One such example is the Swedish verb *be* 'ask, pray', present among the 1,000 words with highest adjusted frequency in SBVP. GUP provides three different lexemes for this verb, either *chiedere*, 'ask', *pregare* 'pray' and *supplicare* 'beseech'. All the words fall into category (C) in Italian, which would not necessarily be true for Swedish. In the opposite direction, the Italian polysemous noun *rapporto*, also among the words in category (C), is covered by three different entries in SBVP, either *förhållande* or *relation* 'relationship', both among the top 1,000 entries, but also 'rapport' 'report', with a lower adjusted frequency.

The degree of coverage of GUP lemmas in SBVP was also measured. It turned out that on overall 37.5% of the translated lemmas were present also in SBVP, but that the (C) group had a significantly higher coverage. Of the total 2,143 candidate lemmas considered as fundamental for communication, 81.4% were also present in the SBVP. Entries in the daily vocabulary (D) group were covered to 20.6%, while 28.6% of the high-frequency lemmas (H) in GUP were present also in SBVP. Of the 483 entries in the Kelly word list which did not occur in GUP, 288 were present in SBVP, i.e. 59.6%.

4.1 The final SweVoc

The GUP and Kelly word list entries that were present in SBVP were used to populate the first four categories in SweVoc, i.e. the core vocabulary items (C), words belonging to every-day language (D), high frequency words (H), and words from modern vocabulary (K). Additionally, items lacking in SBVP but present in both ICF and GUP, denoting daily activities or phenomena, were included. An example of such a word is *andning* 'breathing'. Words present only in ICF, denoting every-day situations and objects, were also added. The Swedish verb *möblera* 'furnish', exemplifies such a word. Finally, a supplementary group of words present only in SBVP were preserved, denoted by the category label (S). The word *samband* 'connection' serves as example from this category. An entry in SweVoc consists of information regarding rank in SBVP, the lemma form, the part-of-speech, and one or more category belongings. The entry *form* is given as example, illustrated below. It is a polysemous noun, found among the 223 most frequent base forms in SBVP, and different senses of the lemma belong to different SweVoc categories.

Rank	**Lemma**	**POS**	**Categories**
223	form	NCU	C, D

In conclusion: the present version of SweVoc contains 7,572 lemmas pertaining to one or more of five different categories. A lemma that is present in more than one category has discriminatory lexical senses, which implies that the number of lexemes amounts to 8,468, see table 2. Category (C) is dominated

Rank	Lemma	Adj.Freq.	Contr.	(WF.PoS.Freq)
5	en.DI	25958.046833	9	ett.DI@NS@S.7952 en.DI@US@S.18050
140	en.MC	726.135618	9	en-.MC0000C.2 ett.MCNSNIS.276 en.MCUSNIS.463
167	en.PI	606.653923	9	ett.PI@NS0@S.147 en.PI@US0@S.467 enom.PIUS0S.1
5708	en.RG	5.661842	4	en.RG0S.9

Table 1: Four different entries of the word *en* in SBVP

by nouns (38%), verbs (23%) and adjectives (13%). The category of words related to every-day matters (D), is mainly composed of nouns (66% of the total amount of lexemes), while verbs and adjectives only occur in 18 and 12% of the totality. In the group of high-frequency words (H), nouns were found to cover 55%, verbs 21% and adjectives 18% of the lexemes. From the perspective of core vocabulary alone, category (C) include 21% of the total nouns in SweVoc, 31% of all verbs, and 23% of all the adjectives. All pronouns and determiners were included in (C). All prepositions except one were found in category (C), except the word *tills* 'until', which was referred to the (D) category. One instance of all conjunctions (*visserligen* 'certainly') was found in the (H) category, while 58% appeared in category (C) and the remaining 40% in category (S). Figures regarding ratios of participles and adverbs are generally somewhat unreliable since different principles were used for corpus part-of-speech tagging in SUC and word list creation of GUP. Specifics regarding the part-of-speech distributions in each category are given in table 3.

Label	Category	Ex	Lexemes
C	Core vocabulary	säga	2,201
D	Words for every-day communication	soffa	1,019
H	High frequency words	sorg	1,518
K	Words in Kelly modern vocabulary	debatt	288
S	Supplementary words from SBVP	ting	3,442
Total			8,468

Table 2: SweVoc entries per category

POS	C	D	H	K	S
Nouns	844	670	831	139	1,436
Verbs	502	181	323	24	575
Adj	295	123	277	52	510
Adv	176	12	8	66	427
Part	168	29	58	7	194
Prep	42	1	0	0	0
Conj	29	0	1	0	20
Pron	65	0	0	0	0
Det	16	0	0	0	0
Other	64	3	20	0	280
Total	2,201	1,019	1,518	288	3,442

Table 3: Part-of-speech distributions in each SweVoc category

5 Evaluation

In order to validate the reliability of the SweVoc, evaluation was performed by coverage tests. It was assumed that the coverage of SweVoc would vary between texts of different types and from various genres. If the core vocabulary items were correctly chosen, the degree of words from this category would correspond to textual complexity, i.e. easier texts would contain more words from category (C). Another assumption was that the ratio of words from category (D) would vary depending on genre, that it would be much smaller, and that the words from the Kelly list (K) would appear more frequently in recent texts. In order to test these hypotheses, evaluation was performed on texts from three different sources:

1. The corpus LäsBarT (LB), which is a corpus of 1.4 million words, containing children's fiction for ages 6-12, and four easy-to-read text varieties:

- Easy-to-read news texts
- Easy-to-read community information texts
- Easy-to-read children's fiction
- Easy-to-read adults' fiction

2. The corpus SUC

3. News text from the daily newspaper Göteborgs-Posten (GP) published in 2007

It was found that, on overall, 91.4% of the tokens in LB, 82.7% of the tokens in SUC, and 83.0% of the tokens in GP were represented at the lemma level in SweVoc, while tokens belonging to the core vocabulary (C) amounted to 80.3% in LB, 68.4% in SUC, and 69.9% in GP texts. The ratios of words related to daily matters (D) were about the same in all texts ($\approx$ 1.3%), but the ratios of high-frequency words were significantly higher in SUC and GP than in LB ($p < 0.001$). Supplementary words (S) were found to be more frequent in SUC than in both GP and LB, which was expected since the original SBVP was retrieved from SUC. By studying the figures in table 4 we can see that the the degree of words in category (C) differ substantially between the ordinary and the easy texts, and also that the percentage of core vocabulary items is higher in fiction than in news and informative texts.

Type/ genre	SweVoc	C	D	H	K	S
ECF	92.5	82.4	0.8	2.1	0.7	6.5
OCF	90.6	80.4	1.0	1.9	0.7	6.6
EAF	93.4	83.1	0.9	2.2	0.6	6.6
OAF	86.3	75.8	1.0	2.4	0.8	6.3
EN	91.5	78.8	1.8	3.9	0.6	6.5
ON	82.2	67.6	1.7	3.8	1.1	8.0
EI	90.6	79.2	1.3	3.3	0.5	6.4

Table 4: SweVoc lemmas, percentage of tokens in different subcorpora
ECF = Children's easy-to-read fiction
OCF = Children's ordinary fiction
EAF = Adults' easy-to-read fiction
OAF = Adults' ordinary fiction (SUC K)
EN = Easy-to-read news
ON = Ordinary news (SUC A and GP)
EI = Easy-to-read community information

6 Foreseen improvements

Entries in the present version of SweVoc preserve information "inherited" from the translated word list GUP, in that a lemma might be categorized with several labels depending on which lexeme it refers to. The Swedish polysemous word *panna* ('front', 'pan', 'oven') is for instance labelled both as a core word (C) and as a word referring to every-day issues (D). One valuable resource for disambiguation is the Swedish word association lexicon Saldo (Borin and Forsberg, 2009), which is a modern Swedish semantic and morphological lexical resource. It is superficially similar to Princeton WordNet (Fellbaum, 1998), but different in the principles by which it is structured. The organizational principles of Saldo consist of two primitive semantic relations, or descriptors, one of which is obligatory and the other optional. When looking up *panna* in Saldo, we find three competing lexemes:

- *panna..1* ansikte..1 ('face') PRIM..1
- *panna..2* laga..2 ('cook') PRIM..2
- *panna..3* elda..1 ('make fire') PRIM..1

The semantic paths in Saldo for each of the three senses are illustrated below, each of the length of 6.

panna..1→ansikte→huvud→kropp→varelse→vem
('face'→'head'→'body'→'being'→'who')
panna..2→laga→mat→äta→leva→vara
('cook'→'food'→'eat'→'live'→'be')
panna..3→elda→eld→brinna→het→varm
('make fire'→'fire'→'burn'→'hot'→'warm')

Frequency counts in SUC reveal that 77% of the instances referred to *panna..1*, 15% to *panna..3*, and 8% to *panna..2*. From these figures, it seems plausible that *panna..1* would be referred to category (C), and either *panna..2* or *panna..3* or possibly both referred to category (D).

Regarding the CALL perspective of this lexical resource, we foresee it as an asset for vocabulary instruction and also as a resource in various CALL-oriented learning platforms and applications, as

for instance the Lextutor, <http://www.lextutor.ca/>. It is also relevant for integration into a Swedish CALL platform under development, cf. Lärka <http://spraakbanken.gu.se/larka/>.

7 Results and conclusion

We found that 81% of the GUP lemmas translated and selected as candidates for inclusion into category (C) were actually to be regarded as pertaining to a core vocabulary for Swedish. Additionally, 21% of the lemmas in category (D) and 29% in category (H) were appropriate for inclusion as complementary vocabulary words.

The resulting word list – SweVoc – of ≈ 7,600 Swedish lemmas is expected to be an asset in language learning and teaching and in readability checkers. The performance of other NLP applications, such as classification tools and morphological analyzers, would also improve with the access of a restricted set of base vocabulary words.

References

Laurie Bauer and Paul Nation. 1993. Word families. *International Journal of Lexicography*, 6(4):253–279.

Lars Borin and Markus Forsberg. 2009. All in the family: A comparison of SALDO and WordNet. In *Proceedings of the Nodalida 2009 Workshop on WordNets and other Lexical Semantic Resources – between Lexical Semantics, Lexicography, Terminology and Formal Ontologies*, Odense.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability. *Educational Research Bulletin*, 27:37–54.

Tullio De Mauro. 1980. *Guida all'uso delle parole*. Editori Riuniti, Roma.

Christiane Fellbaum, editor. 1998. *WordNet: An electronic lexical database*. MIT Press, Cambridge, MA.

Eva Forsbom. 2006. A Swedish Base Vocabulary Pool. In *Swedish Language Technology conference*, Gothenburg.

Dee Gardner. 2007. Validating the construct of word in applied corpus-based vocabulary research: A critical survey. *Applied Linguistics*, 28(2):241–265.

David Hirsh and Paul Nation. 1992. What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language*, 8(2):689–696.

Sofie Johansson Kokkinakis and Elena Volodina. 2011. Corpus-based approaches for the creation of a frequency based vocabulary list in the EU project KELLY issues on reliability, validity and coverage. In *eLex Conference*, Slovenia.

Gunnel Källgren. 1992. SUC - the Stockholm - Umeå Corpus Project: Corpus-based research on models for processing unrestricted swedish text. Technical report, Stockholm.

D. Y. W. Lee. 2001. Defining core vocabulary and tracking its distribution across spoken and written genres. *Journal of English Linguistics*, 29:250–278.

Paul Nation. 1990. *Teaching and learning vocabulary*. Heinle & Heinle, New York.

Paul Nation. 2001. *Learning vocabulary in another language*. Cambridge University Press, Cambridge.

W.W. Patty and W.I. Painter. 1931. Improving our method of selection high-school textbooks. *Journal of Educational Research*, XXIV:23–32, June.

Norbert Schmitt. 2001. *Vocabulary in language teaching*. Cambridge University Press, Cambridge, UK.

Socialstyrelsen. 2003. Klassifikation av funktionstillstånd, funktionshinder och hälsa.

George D. Spache. 1953. A new readability formula for primary-grade reading materials. *Elementary School Journal*, LIII:410–413.

George D. Spache. 1974. *Good reading for poor readers*. Garrard Publishing, Champaign, IL.

Edward L. Thorndike and I. Lorge. 1944. *The teacher's word book of 30,000 words*. Columbia University Press, New York.

Edward L. Thorndike. 1921. *The teacher's word book*. Teacher's College, Columbia University, New York.

M. Vogel and C. Washburne. 1928. An objective method of determining grade placement of children's reading material. *Elementary School Journal*, 28:373–381.

Michael West. 1953. *A General Service List of English Words*. Longman, London.