US-China Foreign Language, ISSN 1539-8080
March 2013, Vol. 11, No. 3, 211-226

Corpus-Based Approaches for the Creation of a Frequency Based Vocabulary List in the EU Project KELLY—Issues on Reliability, Validity, and Coverage

Sofie Johansson Kokkinakis, Elena Volodina
University of Gothenburg, Gothenburg, Sweden

At present there are relatively few vocabulary lists for Swedish describing modern vocabulary as well as being adapted to language learners' needs. In Europe including Sweden there exist approaches to unify ways of working consistently with language learning, one example worth naming in this respect is the CEFR (Common European Framework of Reference) which provides guidelines for systematic approach to language teaching and assessment of language proficiency. This paper describes EU project KELLY (Keywords for Language Learning for Young and adults alike, 2009-2012), the main objective of which was to create vocabulary lists for nine languages (Swedish, English, Norwegian, Greek, Italian, Polish, Arabic, Chinese, and Russian) and adapt them to CEFR levels. We describe the process of compiling and validating the Swedish KELLY-list, dwell on benefits and limitations of using a corpus based approach in this project; as well as mention the impact of the methodological approach for compiling vocabulary lists for specific purposes.

Keywords: corpus linguistics, frequency-based methods, CEFR (Common European Framework of Reference), vocabulary learning, e-resource for Swedish

Introduction

The EU project KELLY (Keywords for Language Learning for Young and adults alike), was granted to ten partner organizations[1] for the period of 2009-2012. The main objective was to develop a bilingual language learning tool for nine languages: Swedish, English, Norwegian, Greek, Italian, Polish, Arabic, Chinese, and Russian, and to adapt it to the above-mentioned CEFR (Common European Framework of Reference) levels[2] (Council of Europe, 2001). Monolingual vocabulary lists for the nine project languages were translated into the eight partner languages, generating 72 language pairs.

CEFR covers six proficiency levels, starting with the beginner level (A1, A2), covering the intermediate level (B1, B2), and up to the mastery level (C1, C2). Proficiency levels are partly defined in terms of what a

---

[*] This paper is a reprint and was originally published in the eLex 2011 proceedings.
Sofie Johansson Kokkinakis, Ph.D., Department of Swedish, University of Gothenburg.
Elena Volodina, Ph.D., Department of Swedish, University of Gothenburg.

[1] Adam Mickiewicz University, Poland; Cambridge Lexicography and Language Services, UK; Consiglio Nazionale delle Ricerche, Italy; Institute for Language and Speech Processing/R.C. "Athena", Greece; Keywords, Sweden; Lexical Computing Ltd, UK; University of Gothenburg, Sweden; University of Leeds, UK; University of Oslo, Norway; and University of Stockholm, Sweden (coordinating partner).

[2] Retrieved from http://www.coe.int/t/dg4/linguistic/cadre_en.asp.

learner should know as far as grammar and communication skills are concerned in the form of "can-do"-statements, and partly in terms of topical (domain) knowledge (e.g., education, sports, etc.). In the light of the above mentioned systematic learning and assessment strategies which are nowadays practiced in Europe and Sweden, the project has been aiming to adapt the selected vocabulary to the CEFR-levels and to evaluate to which extent the CEFR-specific domain vocabulary should be a part of the KELLY-lists.

A corpus based methodological approach was used to ensure that the vocabulary list coverage corresponds to empirically based evidence, and authentic language and language use. The Corpus Factory tool (Kilgarriff, Reddy, Pomikálek, & Avinesh, 2010) was used to aid the creation of a new Swedish corpus (SweWAC (Swedish Web-Acquired Corpus)); SketchEngine (Kilgarriff, Rychly, Smrz, & Tugwell, 2004) was used as a workbench for statistically based selection of potential headwords. Various Swedish electronic lexical resources such as SALDO (Swedish Associative Lexicon) (Borin & Forsberg, 2009) and SMDB (Swedish Morphological Database) (Berg & Cederholm, 2001) were used for proofreading the Swedish list of selected headwords. A database was built to facilitate comparison of the KELLY vocabulary lists and to ensure the validity of the vocabulary item selection across all languages. The final product—a web based language learning tool—is planned to be evaluated quantitatively and qualitatively by a web based vocabulary levels test and a questionnaire.

The Swedish monolingual vocabulary list is at present a freely available electronic resource that reflects a selection of 8,425 most frequent words in modern Swedish as described in Volodina and Johansson Kokkinakis (2012a, 2012b). In this paper we discuss the technologies we have used mentioning their strengths and limitations, and their overall impact on the quality of the Swedish list. Validation and coverage are also described in detail to demonstrate the linguistic appropriateness of this approach.

From Corpus to Wordlist in a Nutshell

The main principle of the KELLY project was that the final vocabulary lists should reflect modern language, constitute the most frequent core vocabulary, plus be based on objective selection.

Corpus Factory for Corpus Collection

To start with, the corpora for vocabulary selection had to reflect present-day language. Moreover, to ensure comparability between vocabulary lists for the nine languages and to guarantee objectivity of word selection, the corpora had to contain at least 100 mln words and preferably to be collected from the web.

Given the above-mentioned prerequisites for the project we faced the problem of an appropriate web corpus of the defined size. There were at the time only two annotated general-language corpora available for Swedish—Parole Corpus and SUC (Stockholm-Umea-Corpus) (Källgren, Gustafson-Capková, & Hartmann, 2006). Neither of the two could qualify as a candidate core corpus for the KELLY-list. Parole dates from 1976-1997 and does not meet the requirement of being a collection of modern language samples. SUC is a balanced corpus dating from 1990's, but comprises only 1.2 mln words and does not meet the requirement of size.

Therefore a big modern corpus of Swedish, a web-corpus SweWAC was compiled by the KELLY partner "Lexical Computing Ltd" using Corpus Factory tool (Kilgarriff et. al., 2010). SweWAC is at present available via commercial concordance tool SketchEngine in its original form (Retrieved from http://www.sketchengine.co.uk/) as well as via the concordance tool Korp freely available through the Swedish Language Bank as a "citation" corpus, in which sentences are mixed in random order so that the full texts

cannot be retrieved (Retrieved from http://spraakbanken.gu.se/korp/).

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 1,000 and 6,000. 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 five 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, and duplicates. The web-corpus finally consisted of 114 million words.

Among the advantages of web-collected corpora we can name the following:

(1) Since its construction is a highly automated process, short collection time at low costs is ensured.

(2) 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.

Among the disadvantages or rather limitations of a web corpus we can name the following:

(1) First of all the absence of control over the kinds of texts which 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.

(2) 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*, and *Daniel*; non-Swedish spelling of words, e.g., *America* (as opposed to the Swedish *Amerika*), British (as opposed to *brittisk*), *company* (Swedish *företag*), *college*, and *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*, and *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), and thus ensure the more or less balanced and predictable structure of the corpus.

(3) Another problem with a web corpus is that automatically collected web-texts may appear with different character encoding.

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 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.

Lemmatization and POS Tagging of SweWAC

Tokenization, lemmatization, and POS-tagging were performed by the Swedish team using the tools

developed by Kokkinakis and Johansson Kokkinakis (1997). SweWAC was tagged for part-of-speech, lemma and morphosyntactic information (case, gender, and number), thus facilitating frequency analysis of word forms, lemmas, and grammatical features. The way lemmatization was performed has naturally influenced the headwords in the KELLY-list. That is why it is important to comment on what we understand by lemma in this context.

The frequency count in the Swedish KELLY-list was calculated upon lemmas (or lem-pos as they are otherwise called), i.e., base form of the word plus its part-of-speech. More closely, in SweWAC context lemma (lem-pos) stands for a set of word forms having the same stem or base form, and belonging to the same word class, e.g., all occurrences of the word forms *flicka*, *flickas*, *flickan*, *flickans*, *flickor*, *flickors*, *flickorna*, and *flickornas* are counted together since they have the same base form *flicka* (Eng. *girl*), the same word class *noun* and the same gender *uter*. However, such definition of a lemma allows grouping together words that share the same base form and word class, but not grammatical features (inflectional morphological aspects), e.g., *fil* (noun, -en, -er; the uter gender, 3rd declension; Eng. *traffic lane*) and *fil* (noun, -en, -ar; the uter gender, 2nd declension; Eng. *file* as in nail-*file*) are counted together in frequency statistics. The missing information about the declension of a noun or conjugation group of a verb results in a partially misleading frequency information. The verb *vara* irrespective of which one of the two verbs is meant—*to be* or *to last*—has always the same frequency value, in spite of the fact that the two verbs are conjugated differently, one being a strong verb (conjugation group 4), the other being a weak verb (conjugation group 1); they also have unrelated meanings, the meaning "*to last*" being much more rarely used. Different lexemes of the same lemma have similarly been summarized, e.g., *rom* (Eng. *caviar*, *rum*, and *Rome*). Thus, neither polysemy nor homography within the same word class have been taken into account during the lemmatization process and consequently during the frequency analysis.

Another aspect which would need further improvement in annotation of the SweWAC corpus is derivational morphology, i.e., mark-up of root morphemes and word-building affixes of each lexical item. The suggested markup could have allowed collecting frequency statistics according to the word family principle, i.e., words that share the same root being grouped together (e.g., *lära*, *verb*, and *lärare*, *noun* would make the same entry). The frequency statistics summarized from SweWAC at present does not allow to group words on this principle, which means a learner that knows the verb *läsa* (Eng. *read*) cannot be assumed to know the noun *läsare* (Eng. *reader*). On the other hand, we do not believe that the word family concept is appropriate for language learners at beginner level.

Errors in frequency calculations of the homographs within the same word class of the type "*vara*, *verb* (Eng. *to be*)—*vara*, *verb* (Eng. *to last*)", though being a systematic drawback, influence only a few rare cases in Swedish and thus have to be neglected in want of a better analysis software. Multiword items that are most frequent in Swedish are marked up as units and do not add misleading information to the statistics used for L2 learners. Finally, taking derivational morphology into account is an arguable demand. Some researchers build their word frequencies upon the notion of word families but they are not many (Gardner, 2007). Thus the two features—having less frequent multiword units, phrasal verbs, and idioms marked up as units, and having roots and affixes marked up for each lemma—refer rather to desirable than to absolutely necessary features. Therefore, we consider word frequency statistics based on lem-pos as described here both reliable and appropriate for language learning purposes.

SketchEngine as a Workbench for Frequency Analysis

To generate frequency-based wordlists over SweWAC, the lemmatized and POS-tagged corpus were uploaded into SketchEngine. SketchEngine offers a number of options for working with statistics. We have used the options of collecting lemma-pos lists with raw frequency alternatively with dispersed frequency.

There are three frequency measures that have been used in the Swedish KELLY-list: RF (raw frequency), relative frequency (word per million or WPM), and ARF (average reduced frequency). 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.

We generated two wordlists: (1) one with lemma-tags in combination with RF; and (2) one with lemma-tags in combination with ARF. The RF-based list from SweWAC contained 402,446 items; whereas ARF-based list contained only 232,900 items. This means that only half of the lemmas in SweWAC have qualified themselves into the general-purpose vocabulary list. We collected raw frequencies for the items on the ARF-ordered list and calculated WPM (word per million) ratio based on raw frequencies. If WPM was less than one, the item was not included into the list. As a result we gained a list of 153,061 items.

Working on Headwords

Word classes. The main guideline in selecting headwords for the Swedish KELLY-list was defined as "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; and avoid proper names with a few exceptions. Homonymy, polysemy, MWE (multiple word expressions), 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, 1,000, 1,000,000, 1st, 2nd, 3rd (but not 4th, 5th), half, quarter, and third.

The following word classes were suggested for exclusion: participle, proper nouns, foreign words, and punctuation.

Prescriptive versus descriptive character. During our work, we came to a point where we had to decide whether our list should be of a prescriptive or descriptive character. On the one hand, the aim of the project was to produce word lists for L2 learners, and in this respect the entries in the list should be of prescriptive character, e.g., incorrect spelling excluded, appropriate words selected. On the other hand, we set as our priority aim to use a modern corpus of Swedish to identify lexical items that are frequent in present-day Swedish, and which therefore are necessary for the language learner to study in the first hand. Thus, if we had started applying

"selection" rules based on our judgment rather than statistics, it would have been a step back and we would risk ending up with a regular list.

On the basis of this, we made a decision to keep our list descriptive in character. That entailed among other things inclusion of several alternative spellings of certain items, and refusal from our part to delete certain vocabulary that did not look "appropriate" for language learners at the pre-translation stage, e.g., words like *stalinistisk*, *adj*, *marxistisk*, *adj*, *sovjetisk*, and *adj* (Eng. *Stalinist*, *adj*; *Marxist*, *adj*; *Soviet*, *adj*). It was planned to check our KELLY-items during the "post-translation" (validation) stage and evaluate every item in the list against translations into Swedish, and if the above-mentioned words could "prove" their basic-vocabulary status by being present in other languages, they would be kept in the final list. If, on the other hand, no other list contained these words, they would be considered for deletion from the final list. Such an approach ensured objectivity and consequence in handling all items, and not only the ones that seemed out-of-place during the initial stage.

Filtering of unwanted words. 30% of 153,061 long list were constituted of "unwanted words and characters" that we removed automatically. By noise we understood the following groups:

All entries (lemmas) containing digits or other characters than letters, e.g., > < = etc.. We preserved items containing underscore (_) since underscores are used in multiword items (e.g., *d_v_s*, *i_alla_fall*).

Some word classes:

Proper names—we have assumed that these were not as important for L2 learners as lexical words. The only proper names that have been added manually to the list are the ones standing for the countries involved in the project (China, Greece, Great Britain, Italy, Norway, Poland, Russia, and Sweden), and large Swedish cities (Stockholm and Gothenburg). Automatic sorting was performed after using a name tagger to differ between nouns and proper names.

Numerals have been removed from the list on the assumption that the number of numerals in the list was too high to handle them manually whereas the most necessary numerals (43% of them) were added manually.

Punctuation marks were removed.

Participles were removed on the assumption that students will learn verbs and eventually learn to apply grammar rules to create participles. Another motivation was that most dictionaries, e.g., SAOL (Swedish Academy Word List), do not provide participles as separate entries; they are, instead, listed together with the verb.

Foreign words which have been recognized by the tagger, were also removed.

Altogether 51,522 lemmas have been removed as "unwanted words" reducing the original 153,061 long list to approximately 100,000 long list.

Final reduction in lemma-number was done automatically by collecting all morphological variants of the same lemma under one unique entry. To illustrate this, the original list contained all forms of the adjective *livlig* (Eng. *lively*):

lemma:-:POStag	Word form
livlig:-:AQPUSNIS	livligt (neutrum)
livlig:-:AQP0PN0S	livliga (plural)
livlig:-:AQPNSNIS	livlig (utrum)
livlig:-:AQC00N0S	livligare (comparative)
livlig:-:AQS00NDS	livligaste (superlative)

All the five forms referring to *livlig*, *adjective* (i.e., livlig:-:AQ) have been reduced to one unique entry for *livlig*, *adj*; all respective frequencies have been summed up resulting in one entry as follows:

ARF	RF	WPM	lemma	POS
572	907.0	7.955	livlig	AQ

The last reduction provided us with a list of 54,338 unique lemmas.

To go through a list of 54,000 lemmas is not an easy task, therefore we cut the list at 9,000-point and started working with it.

Manual analysis of the lemma list. During this stage we made a number of decisions about headwords and the way we want to present them, among other things abbreviations, spelling and form variants, homonymy, polysemy, stylistically marked vocabulary, multiword units, and some marginal cases as described in Volodina and Johansson Kokkinakis (2012b). Lemmatization and tagging errors were identified and fixed, often with the help of concordance searches in SweWAC, for example the noun *fånge* (Eng. *prisoner*) was erroneously lemmatized as a non-existent noun *fångare* from its definite plural form *fångarna*. In some other cases we consulted SAOL online (Retrieved from 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.

Automatic proofreading against other Swedish lexical resources. It is easy, to make omissions during a manual control. Therefore, to double-check that the resulting list contained only existing words, an automatic matching against an associative lexicon, SALDO (Borin & Forsberg, 2009), was performed. About 500 warnings were issued which were double-checked manually—certain passive verbs that did not contain suffix "s" were corrected, e.g., *envisa → envisas* (Eng. *to persist*); some reflexive verbs have been corrected for the reflexive pronoun *sig*, e.g., *befinna → befinna sig* (Eng. *to be present*), some missing word forms in SALDO have proven to be existing in SAOL online; other legitimate items seemed to be too modern to be present in either SAOL or SALDO, e.g., *blogginlägg* (Eng. *blog entry*).

Another automatic control was performed matching SMDB (Berg & Cederholm, 2001), which resulted in a shorter list of warnings which were taken care of manually in the same way as described above.

Finalizing entries for translation. Before sending the list for translation two last steps were performed:

(1) 85 relevant items were added; 43 numerals, 11 geographic names for partner countries, some missing names for family members, words for meals, measures, one missing weekday, and some other domain-specific vocabulary items after comparison with the Swedish Lexicon for Immigrants, LEXIN.

(2) One last manual proofreading was performed where articles were assigned to nouns and infinitive markers to verbs; as well as consistency of headword presentation was checked.

Validation Through Translation

Homonymous and Polysemous Items in Translation

Some teams within the project decided to disambiguate homonymous (and in certain cases polysemous) items manually prior to the translation phase to avoid multiple translations. The Swedish team decided to go after the lem-pos principle to make the process more automatic and fast. It was a part of the decision to run an experiment that will help identify number of one-to-one mappings there are between different language pairs;

number of homonymous and polysemous items which can be identified through translation; and to which extent the list could expand depending on different target languages.

Yet, in certain cases we chose to add an "example" of a typical word context for the translator and eventually for the language learner, though we did not intend to limit the translations by the provided context. We therefore left disambiguation decisions to the subjective judgment of translators.

Translators needed to provide only one translation using the most frequent alternative and to keep in mind that the list was intended for language learners. Where impossible, several translations were provided. The motivation behind the "single translation variant" approach was that items having only one meaning could be used as bidirectional translations of each other, and eventually even multi-directionally between several languages, if translated accordingly. This experiment, demonstrated that this was impossible. If translators had been asked to provide several translation equivalents, it could have secured better mini-lexica. Translation of the polysemous word *rom* provides an illustrative example.

In different contexts headword *rom, noun-en* can mean a drink (Eng. *rum*), food (Eng. *caviar*), a collective name for gypsy people, or a city (*Rome*). In all the cases the noun is of a non-neuter gender, i.e., takes definite ending "-en". Some of the translators showed a "good" sense of humor chose the less frequent meaning of "alcoholic drink" as the most appropriate translation equivalent for L2 learners. Table 1 shows the translation equivalents for the Swedish headword *rom, noun-en* in six languages.

Table 1

Translations of the Swedish Noun "rom"

Language	Translation of the Swedish "rom, n-en"	Meaning in English
English	rum; roe	(1) rum (drink); (2) caviar
Greek	αβγοτάραχο	roe deer
Italian	uova di pesce, rum	(1) caviar; (2) rum (drink)
Norwegian	rom	rum (drink)
Polish	ikra	caviar
Russian	ром	rum (drink)

According to the provided translations, the equivalents for the Swedish *rom* in the other languages are mostly used as a drink, caviar; none of the translators has offered the alternative for the name of the city (probably because of the word class *noun* instead of *proper noun*), nor the collective name for gypsies. The translation also shows that the translated items cannot be used as translations of each other. Generalizing further, we can admit that with the exception of five symmetrically translated items which are mentioned later, none of the translations from the same source word in Swedish can be used as translations between the other 8 partner languages.

Totally there are 2,100 unique Swedish words that have been provided with multiple translations, of those 383 items had multiple translations into more than one language. They were distributed as follows between the CEFR levels: A1-658; A2-167; B1-584; B2-627; C1-497; C2-0 items with multiple translations.

In Table 2, we have collected some information on multiple translations from Swedish per target language.

Table 2
Multiple Translations From Swedish

Language	Multiple translations (homonyms)
English	319
Greek	1,021
Italian	857
Norwegian	1
Polish	325
Russian	7

It is quite unexpected to see only seven multiple translations in Russian that is a more distant relative of Swedish compared to 319 multiple translations into English, a closer language family member. It points to the fact that translation process is highly subjective, and the translator personality and experience influences the resulting work.

The KELLY Database

To make it possible to store, analyze and compare the nine original lists and their translations a special database KELLY DB (database) was created by Lexical Computing Ltd.

Users can search for a word in a web based user interface, and find out whether the word is present in the database and how it is translated into other languages.

Universal, common and unique vocabulary. The main reason for the database was to match original lists for each language with the eight translations into these languages to see how many words are present in all nine languages (symmetric translations, i.e., items that can be safely used as translations of each other), how many are common to eight languages, seven languages, etc., and to generate the following lists: (1) words universal to all nine languages; (2) words specific for each individual language pair; and (3) words unique for each individual language.

A symmetric pair means that the translator of one language, e.g., from English to Swedish has been translated. Let us say *library* as *bibliotek* while the translator from Swedish to English has translated *bibliotek* into *library*. The two translations can therefore be used bi-directionally as translations of each other. A non-symmetric translation can be demonstrated by the following example: (1) *angå* (Swe source item)—*regard* (Eng translation); and (2) *regard* (Eng source)—*betrakta* (Swe translation).

Symmetric set of translations means that (randomly or not) translators between all language pairs chose the same variants for the pairs "source word"—"target word".

It has turned out that only five words belong to the universal vocabulary, i.e., they are translated in symmetrical sets. These words are *music*, *library*, *sun*, *hospital*, and *theory*. The constellation of the "universal" vocabulary appears to be rather random depending on translators' preferences, and seems to rely on chance rather than on some linguistic reasons.

Surprisingly enough some expected words like weekdays, months, numbers, and names for relatives and basic foods have not gained the status of universal vocabulary. For example, the word *bread* is (almost) symmetrically translated, with the exception of one translation where an extra variant (synonym)—*corn*—is provided. The same refers to the word *mother*: All translators into Swedish chose the variant *mor* except the one who translated it with *moder*. As far as *father* is concerned, there were different translation variants to Swedish, including *pappa*, *far*, and *fader* which made translation sets asymmetrical.

The symmetrical sets for eight and seven languages do not seem to reveal much of a language either apart from the fact that certain languages have more variants for the same notion, and therefore they do not add to the symmetry. Certain asymmetrical sets are the result of incorrect translations or different interpretation of the source words. A very interesting example is weekdays. In Chinese at least three different names for each weekday are used (depending on the translation equivalent for *week*). In Arabic there are at least two names for each weekday; which of course has made it impossible for weekdays to enter a symmetric set for nine or eight languages.

Absence of ordinal numerals (one, two, and three, etc.) among symmetric sets for nine or eight languages is also rather surprising at first glance. It takes to know the other languages to see the reason why it happens that way.

The numbers for common vocabulary between different language pairs comprise symmetric pairs for each language combination. Table 3 shows the numbers for languages paired with Swedish.

Table 3

Common Vocabulary for Swedish-X Language Combinations

Language combination	Nr of symmetric pairs
Swedish-Norwegian	3,109
Swedish-English	3,002
Swedish-Italian	2,641
Swedish-Polish	2,495
Swedish-Russian	2,271
Swedish-Greek	1,966
Swedish-Chinese	1,123
Swedish-Arabic	618

The numbers indicate how many entries in the two languages can be used bi-directionally.

Numbers of the common vocabulary between different language pairs seem to confirm the fact of "closeness" between the languages depending on which language family they belong to—the closer relatives the languages are, the more common vocabulary (symmetric pairs) they share. It also reflects relative similarity of the corpora from which the original lists have been derived as well as approaches to vocabulary selection.

The highest number of symmetric sets enjoys the pair Swedish-Norwegian: Both languages belong to the same language family, subgroup and branch (Indo-European family, Germanic Subgroup, and Northern branch). Both lists have been derived from web corpora. Swedish-English pair comes next. Both these languages belong to the same family and subgroup, the difference lies in the branch (Northern vs. Western). English list has been derived on a combination of different corpora since there are many more available for English than for Swedish.

The least number of symmetric pairs is shared by Swedish and Arabic, which reflects distance between languages (Germanic vs. Afro-Asiatic language families) and the principles of tokenization, lemmatization and vocabulary selection.

Unique vocabulary in this context means the items present in the monolingual list that were not used in any of the translations from other languages to the target language.

There are 501 words in the list of unique Swedish words. They represent 118 words marked for domains, while 370 come from the "exclusion list". The latter ones are kept for the reasons described later, among them are Swedish-specific words like *midsommar*, *pingst*, *nobelpris*, *kvällsmål*, and *fika* (Eng. *Midsummer*, *Treenity*, *Nobel Prize*, *supper*, and *coffee break*).

The lists of universal, common and unique vocabulary may present certain interest for lexicographers, comparative linguists and other language-interested user groups and have a potential for being further exploited in linguistic analyses. The Swedish list is available for download at the Swedish Language Bank.

Inclusion and exclusion candidate lists. Apart from that, the KELLY database facilitated generation of the following lists necessary for post-translation editing and validation of the monolingual master lists (M):

(1) Candidates for exclusion for each individual language, i.e., words present in the target monolingual list but not used in any of the translations from other languages to the target language; (2) Candidates for inclusion, i.e., words that have been used as translations to the target language, but are not present in the target language monolingual list; and (3) Multiword expressions not present in the original monolingual list, but given as translations into the target language from other languages.

Embedding the Evidence

The Swedish M2 (Monolingual, 2nd revision) list sent for translation contained 6,000 items. After processing the candidate lists generated from the KELLY DB, it expanded to 8,425 items. This confirmed our intuitions that translations from other languages could enrich each language with approximately 2,000-3,000 items.

The deletion candidate list for Swedish contained 644 candidates for exclusion, i.e., 644 lemmas that have not been used as translations into Swedish from any of the eight partner languages. We went through the deletion candidates manually, deleted 137 items from the monolingual list and kept 507, guided by the principles described in Volodina and Johansson Kokkinakis (2012a), the most important one being the domain of importance to language learners, e.g., *veckodag* (Eng. *weekday*), *vänster om* (Eng. *to the left of*), and culturally important words for Swedish, e.g., *midsommar* (Eng. *midsummer holiday*), *fika* (Eng. *coffee break*).

We deleted items from the Swedish M2 list if the deletion candidates were words that had functional word classes, e.g., particles, determiners, and pronouns; historical terms, e.g., *stalinistisk*, *bolsjevik*, *marxistisk*, and *koncentrationsläger*; adverbs if they were "t"-derived forms from an adjective present in the M2 list, and some other groups as described in Volodina and Johansson Kokkinakis (2012a).

Inclusion candidates list comprised 3,430 base forms. Of those, 2,630 lem-pos have been added. The 3,430 candidates were first automatically checked against a SweWAC lemma list, and all possible POS-tags for each item and their WPM frequencies were collected. A number of items did not match any of the lemmas in the SweWAC and were discarded as illegitimate ones. Among the latter ones there were non-lemmatized items, e.g., *dikter* (Eng. *poems*), non-existent or misspelled word forms.

Due to the collected SweWAC wpm frequencies, it was possible to place all inclusion candidates relative to the items already on the Swedish list. Most of the added candidates ended up in the last two proficiency levels on the Swedish list.

Out of 530 candidate MWE, examples (as opposed to headwords) were added to 115 headwords, to 44 of those multiple examples. Altogether 194 MWE were added to the list. We discarded non-idiomatic and non-lemmatized candidate MWEs, e.g., *bära in* (Eng. *bring in*), *bära ut* (Eng. *take out*). We avoided inclusion of MWE as new headwords since we did not have the frequency for those.

As for the lacking domain specific vocabulary, only frequency justified topical words from the eight languages were added in the Swedish list, thus making the selection of domain vocabulary also based on the frequency principle.

Coverage

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 in the following way: (1) A1, A2, B1, B2, C1—1404 headwords per level; and (2) C2—1405 headwords.

With respect to their sources, the headwords are distributed in the following way:

(1) 85 have been added manually. They constitute 1% of the list, all belonging to CEFR A1 and cover 0.44% of SweWAC.

(2) 2,564 headwords come from T (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 4.

(3) 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 (between 1,305 and 1,377 headwords per level) in the first four CEFR levels, and disappear at all from the last CEFR level C2, as shown in Table 4.

Table 4

SweWAC Coverage by T2 and SweWAC Items

CEFR level	Nr of T2 words	SweWAC coverage (%)	Nr of SweWAC items	SweWAC coverage (%)
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)	1,405	0.2476	0	0
Total	2,564	1.6739	5,776	77.98

Word class distribution is presented in Table 5.

Table 5

KELLY POS Distribution in SweWAC

POS	Total count (% of KELLY-list)	Coverage, SweWAC (%)
Adjective	1,354 (16.07)	6.43
Adverb	569 (6.75)	7.6
Aux. verb	5 (0.06)	0.14
Conjunction	19 (0.23)	0.41
Determiner	10 (0.12)	3.6
Interjection	24 (0.28)	0.1
Noun	4,607 (54.68)	14.51
Numeral	56 (0.66)	1.19
Participle	1 (0.01)	0.001
Particle	29 (0.34)	0.45
Preposition	108 (1.28)	11.14
Pronoun	61 (0.72)	11.4
Proper name	13 (0.15)	0.24
Subjunction	31 (0.37)	1.8
Verb	1,538 (18.26)	16.9

61 pronouns covered 11.4% of SweWAC; 108 prepositions covered 11.14%; whereas 4,607 nouns covered only 14.51% compared to 1,538 verbs which covered 16.9%. Verbs, pronouns, and prepositions therefore appears more "beneficial" to learn than of nouns in terms of text coverage, or so it would seem from statistics.

Corpora Coverage by KELLY-Items

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 in nine genres. SUC has been manually proofread for errors in lemmatization and part-of-speech tagging.

Coverage calculations indicates that words from the Swedish KELLY-list cover 80% of the total of SweWAC, punctuation, infinitive markers and proper names stand for 16%. However, coverage calculations of the two other corpora have shown that KELLY words cover only 62.75% of the Parole corpus and 68.87% of the SUC corpus as illustrated (see Table 6).

Table 6

SweWAC, Parole, and SUC Coverage in %

Parameter	SweWAC	Parole	SUC
Size	114 mln	25.7 mln	1.16 mln
Language	2010's	1976-1997	1990's
Type of corpus	web-acquired	general-purpose (written) language	general-purpose (written) language
Annotation (POS, lemma)	Yes	Yes	Yes
Punctuation	10.7%	12.7%	11.5%
Infinitive marker	1.26%	1.01%	1.1%
Proper names	4.87%	8.67%	3.6%
KELLY-words	79.65%	62.75%	68.87%
Total coverage	96.5%	85.14%	85.07%

A number of KELLY-items got zero-matches in the control corpora: 653 items did not appear at all in SUC and 224 had no match in Parole. Reasons might be: (1) differences in tagging and lemmatization; and (2) difference in text genres constituting the three corpora.

(1) The first difference lies in tagging and lemmatization. Lemmatization and pos-tagging of the two control corpora differ from the SweWAC-based KELLY-list. Even though Parole was tagged and lemmatized the same way as SweWAC, the headwords in the KELLY-list have undergone manually introduced changes. As a result a number of items were corrected for word class tags or lemma, for example *själv* (Eng *self*) changed pos from *adjective* to *pronoun* in the KELLY-list. In Parole *själv* is alternatively tagged (in certain cases erroneously!) as *adjective*, *noun* or *adverb*. Tagging differences can also be seen in POS-mismatches in such highly frequent words as *ett*, *det*, *sin*, and *annan*, etc., that are tagged as *pronouns* in the KELLY-list as opposed to *determiner* in SUC.

A number of headwords in the KELLY-list have been modified to make them more user-friendly for L2 learners. For example, the reflexive verb *te sig* had originally been lemmatized and POS-tagged as *te*, *verb*, but was manually corrected during the work on the KELLY-list to *te_sig*, *verb*. Thus, none of the lemmas in Parole matched the KELLY-item *te_sig*, nor any other reflexive verbs for that matter. Generally, verbs appearing among zero-matches fall into two categories: the above-mentioned group of reflexive verbs (e.g., *te_sig*), and -s verbs that originally have been lemmatized without the final "-s", but have been manually corrected in the KELLY-list, e.g., *vista* vs. *vistas* (Eng. *to stay*).

A big group of POS-mismatches are items tagged as *adjectives* in the KELLY-list, while having *participle* tag in SUC and Parole, among them *nuvarande*, *anställd*, and *växande* (Eng. *present*, *employed*, *growing*).

Some multiword expressions have been manually corrected by us in the KELLY-list and did not find any correspondences in either Parole or SUC, e.g., *till_slut*, *på_sistone*, *i_närheten_av*, *varken… eller* (Eng. *in the end*, *of late*, *in the vicinity of*, *either… or*).

(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 vocabulary development of the recent decade:

(a) The zero-matches reflect recent "hot" political events and technological innovations, e.g., *piratparti*, *svininfluensa*, *alliansregering*, *islamist*, *taliban*, *reporänta*, *fildelare*, and *sms* (Eng. *pirate party*, *swine flu*, *alliance government*, *Islamist*, *Taliban*, *funding rate*, *file sharer*, and *sms*);

(b) The zero-matches make it obvious that the domain of web-related texts and computer technologies dominate in SweWAC, e.g., *blogga*, *bloggare*, *blogginlägg*, *textstorlek*, *postning*, *webbläsare*, and *webbsida*, (Eng. *to blog*, *a blogger*, *blog entry*, *font size*, *posting*, *web browser*, and *website*);

(c) Some other vocabulary absent in SUC and/or Parole is very colloquial in its nature and can be taken as evidence of more colloquial character of online conversation that constitute a part of SweWAC (blogs, chats, and forums), e.g., *toppen*, *jävla*, and *tryne* (Eng. *great*, *damn*, and *snout*);

(d) Absence of down-to-earth learner-specific domain vocabulary in SUC can be demonstrated by the words coming to KELLY-list from translation lists, such as *krabba*, *socka*, *huva*, *sparv*, *sesam*, *aprikos*, and *brorsdotter* (Eng. *crab*, *sock*, *hood*, *sparrow*, *sesame*, *apricot*, and *niece*)

(e) One more group of zero-matches is constituted by widely spread loaned words such as *shopping*, *klick*, *mejl*, *kidnappning*, *designer*, and *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 reason (2) above is both modern and relevant vocabulary for L2 learners.

Conclusions

Time Aspect

The linguistics part of the project described included generation of mono- and bilingual lists during a period of 4 months of full-time work for the Swedish team. The five-step process for generation of the Swedish list took time as shown below: (1) corpus creation and tagging—2 months; (2) frequency lists generation via SketchEngine—1.5 weeks full-time work; (3) working on headwords—6 weeks full-time work; (4) translation—4 months; and (5) validation—7 weeks full-time work.

Using automatic methods is necessary when dealing with large corpora, but some automatic processes are not fully satisfactory, e.g., lemmatization, identification of multiword expressions, phrasal verbs and lexeme differentiation into the first version of the frequency list.

Various types of error correction of the first version of the vocabulary list was time consuming but necessary.

The Source Corpus

The process of creating learner-oriented word lists should start with a well composed and balanced corpus. The best approach is to use some available balanced representative corpus of modern language that is large enough for the task. If such corpus is not available, the web-corpus is the best and fastest alternative, though in that case we suggest that the language team be asked to provide a list of seed words. It is then possible to "design" a balanced web-corpus with seed words selected for different genres. The list of genres can be complemented as necessary; seed words for each genre carefully preselected manually or generated automatically from a shorter existing balanced corpus that contains a number of genres. Genre corpus will presumably prevent obvious gaps in learner-specific domain vocabulary, e.g., lack of words like *orange*, *elbow*, or *alphabet*.

Multiword Expressions and Lexeme Differentiation

Phrasal verbs, idioms and multiword expressions are definitely valuable items on any list, to say nothing of the learner-oriented lists. The question is whether existing NLP (Natural Language Processing) tools display sufficient accuracy.

As far as word sense disambiguation and lexeme-based frequency calculations are concerned, we are back to the fact that there are no reliable tools for Swedish at the moment that can either disambiguate word senses and collect frequency statistics per lexeme, or differentiate between homography within the same word class with sufficient accuracy. However, we can hypothesize that having the same lem-pos several times in the list in different proficiency levels (i.e., homographs or different lexemes) might be confusing for a language learner. A learner who identifies a token "sentence" in a text and who has for the reason of frequencies learned only one meaning of this token, let us say within the domain of linguistic meta-language, will be baffled when he sees the item in the "legal" context: *He had his prison sentence reduced*. It is probably better to inform the learner of other possible meanings of the lem-pos the first time they come across it, so that they know they need to go back to that item and check additional meanings when they encounter it in an unknown context.

Future Plans

We can conclude by saying that we plan to continue working with the Swedish KELLY-list. The way it has been extracted, it addresses a number of target user groups, including language teachers, test creators, lexicographers, comparative linguists, and computational linguists, etc.. In the near future we plan to set up a dynamic lexical database where different types of word lists can be extracted, e.g., items per domain, per CEFR-level, items shared by different language pairs, words that have received multiple translations etc.. The users will be able to add corpora examples and translations to the items in a dynamic way. Linking this database to other lexical resources available through the Swedish Language Bank (spraakbanken.gu.se), the intention is to provide for automatic analysis of morphological constituents of each item and experiment with other interesting options.

Another path we want to pursue is within language teaching, among other things we plan to test how many words learners of different CEFR levels know; whether the words are assigned to the appropriate CEFR-levels; and run coverage tests on language course text books used in CEFR-based language courses.

References

Berg, S., & Cederholm, Y. (2001). On the creation of the Swedish morphological database (Att hålla på formerna: Om framväxten av Svensk morfologisk databas). In *Concerning stems, suffixes and words (Gäller stam, suffix och ord)* (pp. 58-69). (Publication in honor of Martin Gellerstam, October 15, 2001). Mejerberg's archives for Swedish vocabulary research (Meijerbergs arkiv för svensk ordforskning 29). Gothenburg: Elander Novum.

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.*

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

Kilgarriff, A., Reddy, S., Pomikálek, J., & Avinesh, P. V. S. (2010). A corpus factory for many languages. *Proceedings of LREC*. Malta. Retrieved from http://www.lrec-conf.org/proceedings/lrec2010/

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). The sketch engine. *Proceedings of EURALEX*. Lorient, France: Retrieved from http://www.euralex.org/elx_proceedings/Euralex2004/

Kokkinakis, D., & Johansson Kokkinakis, S. (1997). A robust and modularized lemmatizer/tagger for Swedish based on large lexical resources (Swedish Language Department, University of Gothenburg).

Källgren, G., Gustafson-Capková, S., & Hartmann, B. (2006). *Manual of the Stockholm Umea corpus* (p. 85). Department of Linguistics, Stockholm University.

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

Volodina, E., & Johansson Kokkinakis, S. (2012a). *Swedish Kelly: Technical report*. University of Gothenburg: The Swedish Language Bank.

Volodina, E., & Johansson Kokkinakis, S. (2012b). Introducing Swedish Kelly-list, a new free e-resource for Swedish. *Proceedings of LREC*, 2012, Turkey.