GÖTEBORGS UNIVERSITET
INST FÖR SVENSKA SPRÅKET

GU-ISS-2012-01

Swedish KELLY: Technical report

Elena Volodina
Sofie Johansson Kokkinakis

Forskningsrapporter från institutionen för svenska språket, Göteborgs universitet
Research Reports from the Department of Swedish

ISSN 1401-5919

www.svenska.gu.se/publikationer/GU-ISS

Contents

Introduction

KELLY is a European Union project funded by the EUs Lifelong Learning Programme, KA2 Languages subprogramme. It was granted in 2009 to 10 partner organizations:

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

Keewords, Sweden

Lexical Computing Ltd, UK

University of Gothenburg, Sweden

University of Leeds, UK

University of Oslo, Norway

University of Stockholm, Sweden (coordinating partner)

The project was financed for two years starting on 01-11-2009.

KELLY stands for the shortening of KEywords for Language Learning for Young and adults alike, the name itself reflecting the main aim of the project – identifying keywords in a language for language learners. More precisely, we set out to identify approximately 9000 most frequent words for a language corresponding to the European Framework’s six study levels, plus to develop a language learning product with the above-mentioned words and their equivalents in another partner language to promote vocabulary learning.

There are 9 partner languages that are involved in the project: Arabic, Chinese, English, Greek, Italian, Norwegian, Polish, and Swedish, which means that by the end of the project bilingual lists with 72 language pairs were prepared (e.g. English-Chinese, Chinese-English, Swedish-Norwegian, Norwegian-Swedish, etc.). These bilingual lists are supposed to function as a basis for complementary learning material, the target group being language learners of these nine languages, 16 years and up, that study a language in upper secondary school, evening school, or at a university.

Structure of the report

Work on the project was divided into several Work Packages including management, dissemination, linguistic analysis, evaluation, production of the learning tool, quality plan and exploitation of results. In this report only Work Package 3 ”Linguistic Analysis” is described.

During this work package the partners were supposed to:

produce frequency lists based on a 100-mln-word corpus, cut at 6000 words;

clean up and proof-read the lists;

send the lists to the translation agency for translation into 8 partner languages;

merge each original word list with the 8 translations from other languages;

finalize the lists by checking/adding candidates for inclusion/exclusion plus evaluate the necessity to add specific domain vocabulary important for the language learners that might be absent in the lists

Chapter 2 is devoted to the pre-translation phase of the work package 3, where we describe the workflow, decisions we have made, problems we have identified and lessons we have learned.

Chapter 3 describes the modifications made to the Swedish KELLY-list during the post-translation phase, including problems, decisions and lessons learned. Some analysis of the results of translation is provided.

In Chapter 4 we provide information on Kelly database (Kelly DB) and the first experiments with the coverage by the Swedish Kelly list.

Prior to chapter 2 we felt it was necessary to provide a short description of European Framework study levels referred to earlier and to make a short summary of available word lists for Swedish aimed at second/foreign language learners to show the reader how the KELLY-list differs from the existing lists and what advantages it has.

Common European Framework of References for Languages (CEFR)

CEFR is a document containing guidelines for language teaching and for ascribing proficiency levels to learners of European languages, and of late, borrowed even to some non-European languages. The initiative to harmonize language learning levels across countries was raised in 1991 in Switzerland, the work on level descriptions being finished by 1996. The language assessment scale contains 6 levels:

A Basic Speaker
- **A1** Breakthrough
- **A2** Waystage

B Independent Speaker
- **B1** Threshold
- **B2** Vantage

C Proficient Speaker
- **C1** Effective Operational Proficiency
- **C2** Mastery

The language proficiency levels are described in the form of can-do statements[1]:

Level	Description
A1	Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.
A2	Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can

[1] Source of information:
<http://en.wikipedia.org/wiki/Common_European_Framework_of_Reference_for_Languages>

	communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.
B1	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes & ambitions and briefly give reasons and explanations for opinions and plans.
B2	Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.
C1	Can understand a wide range of demanding, longer texts, and recognise implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices.
C2	Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in the most complex situations.

CEFR is said to be used as a reference frame and should be adjusted locally within each country. Many countries with a long tradition of ascribing other proficiency levels to the language learners have preferred to abandon their local assessment standards in favor of CEFR, for example Sweden. As an illustration, the national Test In Swedish for University Students (TISUS) that used to give “svenska B” level now is announced to give level C1 according to CEFR.

Attempts have been made to identify how many hours each level can demand in teacher-driven education (Deutche Welle). However, up to date there has never been any description for Swedish of what exact vocabulary learners of each CEFR level should master, or how many words on each level. That is where Swedish KELLY-lists for 9 language combinations come in handy.

General on vocabulary learning and on the use of frequency-based wordlists

Words are recognized as essential building blocks of the language. Language users that know the grammar of a language cannot explain themselves if they do not know words. However, knowing words without knowledge of grammar can

help communicate ideas. Lexical competence is therefore important for language acquisition and effective communication.

Native speakers develop their lexical competence in early childhood, filling the existing blanks in response to new experiences as the need arises, i.e. incidentally. For second language learners the picture is more complicated: vocabulary acquisition is a conscious and time-consuming process that has to be supported by specially designed activities for more effective progress.

Vocabulary can be acquired in different ways – through conscious learning (e.g. memorizing lists of words, doing vocabulary exercises, using target vocabulary in speech or writing) or through incidental learning (e.g. reading, listening). The fact remains though: vocabulary acquisition should be assisted if the learner is to develop good lexical competence in a fast and effective way (Nation & Waring 1997; Read 2000; Ma & Kelly 2006).

To stimulate better vocabulary learning appropriate words for each learner level should be selected. On what grounds should this vocabulary be selected? How should it be divided into levels? Are there any general recommendations? How do teachers identify those words in their everyday practice?

These questions are often asked to the Swedish Language Bank (Språkbanken), which in itself says a lot about the need for such guidance. The general recommendations that we usually give are to use some of the resources listed in section 1.4. It is, however, not a totally satisfactory answer since neither of the lists below can offer modern language in combination with streaming into either difficulty levels or frequency bands.

We have turned to different organizations in Sweden that have responsibility for education with the same questions. Among those were Swedish Language Council (Språkrådet), Ministry of Education (Utbildningsdepatrtementet) and people responsible for TISUS (Test In Swedish for University Studies). Neither of these has provided us with any information on available modern Swedish word lists based on frequency statistics and streamed into difficulty levels. Swedish Language Council expressed interest in the prospective KELLY-list for Swedish for future use.

Available wordlists for Swedish language learners

The information that follows below includes a short summary and publisher details of vocabulary resources aimed at Swedish language learners available today.

1. The special learner dictionary "Natur och Kulturs Svenska Ordbok" (="Swedish Words") published by "Natur och Kultur" (Köhler & Messelius, 2006) contains 23.000 words + 9.000 idioms and set phrases that represent Swedish central vocabulary necessary for learners. The selection is claimed to be done based on other dictionaries and frequency studies, plus certain personal judgment of the people involved in the project, e.g. all words related to nationalities have been removed and placed as an appendix, some other learner relevant items added etc. Entries do not have any frequency information and it is uncertain how this source can be used for selecting appropriate vocabulary for different study levels. A noteworthy feature of the dictionary is articles before nouns plus an exercise book available for copying.

We are aware of the paper version only. ISBN 9789127570627

2. A special learner wordlist "Svensk skolordlista" (="Swedish wordlist for schools"), 35.000 items, published by Norstedt (Nygren 2010), is a wordlist that has been prepared in collaboration between the Swedish Academy (Svenska Akademien) and the Swedish Language Board (Svenska språknämnden). It is aimed at pupils from the 5th grade and up, and contains short explanations in easy Swedish to almost all the items on the list.
This list is based on the SAOL (Swedish Academy's Word List of Swedish Language, updated regularly, approx 125.000 words). The selection of 35.000 words has been made on the basis of most frequent words in modern newspapers and books, including a number of colloquial words (used in speech rather than written texts), plus somewhat outdated words typical of literature/fiction used for learners of Swedish. No frequency information provided so it is not clear how words can be streamed into difficulty levels.
Paper copy: ISBN 9113028529. We are not aware of the electronic version.

3. " Svenska ord: med uttal och förklaringar" (Lexin 2006) is a dictionary available both as a paper copy and a web-based dictionary. In English its title says "Lexin Swedish words with pronunciation and explanations". The paper copy has been released in 2005 (3rd edition); the web-based version has been updated in 2011. This dictionary contains 28.500 words and is aimed at immigrants as a target group. The vocabulary has been selected according to the following:
- Swedish central vocabulary comes from frequency studies (no details) plus Sture Allén's "Våra viktiga ord" (2002) (see description below);
- Vocabulary collected from course books for immigrants, e.g "Svenska för invandrare" ("Swedish for immigrants"-series);
- Words specific for social studies (samhällsord) partly manually selected and partly coming from specific interpreter lists;
- Colloquial words and "difficult"-for-learners words come from about 20 different sources, that are described in Gellerstams "Välja sina ord" (1978).
This dictionary is regularly updated based on corpus studies; certain vocabulary is added/removed following tests on words carried out in schools, comparing native learners versus Swedish language learners. Yet, there is no frequency information; neither information on the vocabulary appropriateness for different learner difficulty levels.

This dictionary includes a topical picture section for some most important areas.
ISBN 978-91-85128-58-7 / 91-85128-58-9
online dictionary: http://lexin.nada.kth.se/lexin/

4. The frequency lexicon "Tiotusen i topp" (="Top ten thousand") by Sture Allén (1972) is a frequency list of 10.000 most frequent words in Swedish. The list has been produced on the basis of newspaper texts collected around 1965 and is claimed to be very useful in education. Distribution/normalization has not been taken into account. The book contains 6 parts:
- 10.000 most frequent graphical forms in frequency order (incl. freq info);
- 10.000 most frequent lemmas in alphabetic order per each thousand (not incl. freq info);
- 10.000 most frequent lemmas in alphabetic order per part of speech (incl. freq info);
- 10.000 most frequent words in final/backward alphabetic order (ordered after the last letters, useful for rhymes and crosswords) (incl. freq info);
- 10.000 most frequent lemmas in alphabetic order (incl. freq info);
- letters and other characters in their frequency order.
This list has two drawbacks: it has never been updated since 1972 and it does

not take into consideration dispersion.
Interesting information is that the words from the first thousand cover 70% of the whole newspaper corpus; all of 10.000 words make up 90% of the newspaper corpus.
Paper copy: ISBN 9120051840

5. Base Vocabulary Pool by Eva Forsbom (2006) is a frequency based word list constituting central vocabulary derived from the SUC (Stockholm Umeå Corpus). The base vocabulary pool is created on the assumption that domain- or genre-specific words should not be the basis of a base vocabulary pool. The core of this list is constituted by stylistically neutral general-purpose words collected from as many domains and genres as possible. As a result out of 69,371 entries in the lemma list based on SUC, only 8,215 lemmas have qualified themselves into the base vocabulary pool, and they account for 88.2% of all the SUC texts.

Base Vocabulary Pool is a very good resource but a bit short for our purposes. SUC that has been used as a source corpus dates from 1990-s, contains mostly written texts and has 1,2 million running words.

This list is publicly available in electronic form from <http://stp.lingfil.uu.se/~evafo/resources/basevocpool/> (under the heading “Files”, data -/base vocabulary pools, “SUC_basevoc”)

6. "Våra viktiga ord" (="Our most important words") by Sture Allén (2002) is a dictionary that explains approx 7.000 basic vocabulary words that are specifically important for learners of Swedish. These words have been selected on the basis of frequency information and validated against two other word lists prepared by teachers. The final list is the result of merging the three source lists. The dictionary entries include base form of the word, morphologically conjugated forms and pronunciation where it is necessary. This resource is too short for our aims and does not contain frequency information for streaming vocabulary into difficulty levels.
ISBN10: 9121199701
ISBN13: 9789121199701

7. "Libers lilla ordlista" (="Liber's little wordlist") by Sture Allén (2006) is also a dictionary explaining 8000 central words. The dictionary contains a picture section with some important groups of words, e.g. “In a classroom”. There is a student book for training vocabulary. It contains neither frequency information, nor the information on principles for selecting central vocabulary.
ISBN10: 9147081341
ISBN13: 9789147081349

8. “Praktisk Svensk Ordlista” (=”Practical list of Swedish”) (1993) was published by Swedish Academy and Swedish Language Council. It is based on SAOL and contains 30.000 entries. Selection of words for inclusion is not clear, it is mentioned that the most important words have been selected, including some vulgar words. Compound words that have a clear meaning deducible from the stems have been excluded. Some foreign words have been taken to demonstrate that there are Swedish alternatives that can be used instead.

Each word is provided with a short definition, conjugated forms, and some other information. The target group for this dictionary is not specified in the introduction.

There is no frequency information accompanying individual words.

ISBN 91-1-935372-3

2. Pre-translation phase

The main principle of the KELLY lists was that they should reflect the modern language, constitute the most frequent core vocabulary, plus be based on objective selection (avoiding human judgment as much as possible). Besides, vocabulary should be streamed into CEFR difficulty levels.

This interprets into the following:

The corpora that the vocabulary selection is based upon should be samples of present-day language. Moreover, to ensure comparability between word lists for the 9 partner languages and to guarantee objectivity of word selection, the corpora should contain at least 100 mln words and be preferably collected from the web.

1. To ensure that only domain-free language comes into the frequency list a special “weighting” of each word should be carried out, which means that each word has to be checked whether it is frequent in a few texts of a certain domain (e.g. law or medicine) or it is regularly used in all types of texts. There are several methods to check that automatically, the one that has been used by our team is average reduced frequency (ARF) as described in Savický & Hlavácová (2002).
2. The word selection should be strictly frequency-based. All pedagogical “modifications, additions and deletions” should follow straightforward principles and be reproducible in case someone will want to repeat this experiment.
3. Streaming into language levels (number of words in each level plus some domain-specific vocabulary necessary for language learner per CEFR level as mentioned in (3)) should follow the frequency principle or some other objectively-defined method.

As can be seen from the description above, neither of the available word lists for Swedish described in 1.4 matches the requirements set on KELLY word list. The way the KELLY list is compiled, it should be a reliable resource for defining a 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 uses.

The linguistic phase of this project has been given 10 months, from February, 2010 till December 2010; the pre-translation phase being given the period of 3 months, translation phase – 4 months and post-translation phase - 3 months according to the “Action Plan from Athens Meeting” (available on the project wiki-page).

As a result of the Athens meeting the following workflow has been defined for producing the first version of monolingual word lists (M1 -> M2) for each language:

1. Identify the core and the reference corpora
2. Lemmatize and POS-tag both corpora using the same tools
3. Compare core and reference corpora and integrate evidence from the reference corpus
4. Generate a lemgram list (lemgram = lemma plus its part of speech, POS) from the core corpus taking into consideration dispersion, e.g. create lists based on ARF (average reduced frequency) using SketchEngine, if possible (M1)

5. Edit the core word list (M1) with respect to different linguistic aspects if appropriate for a language and deliver a second version of the list (M2):
a. Filter words with other characters than a-z, e.g. containing numbers
b. Filter proper names e.g. for English “anything capitalized isn’t core vocab, exclude unless covered by a special case” (from “Action Plan from Athens Meeting”)
c. Merge spelling variants
d. Take decision and actions thereafter on marginal classes (numerals, prefixes, days of week, etc.)
e. Take decisions on homonymy, polysemy, etc. and take actions thereafter
f. Edit obvious multiword expressions
6. Prepare spreadsheet that includes certain columns for translators, add translation instructions and mail the resulting document(s) to the partner responsible for “subcontracting” translators.
   The description below follows the steps described in the workflow.

2.1 Corpora availability for Swedish

In pre-corpora times language teaching materials have been selected based on the intuition of course-book writers and/or teachers. Now that corpora are available it is possible to check those intuitions by consulting automatically generated frequency lists over different features tagged in a corpus and make conclusions about which features are most typical, e.g. most frequent and presumably most important for language learners. Some teacher intuitions referred to above can be confirmed right, others – proved wrong. For instance some language teachers working with corpora have come to an insight that certain language course books tend to overestimate importance of the verbs “will” and “shall” as expressions of future in English overlooking the fact that native speakers prioritize other ways of expressing future.

It is also true that frequency alone cannot be the only factor for consideration when it comes to learner material selection. For example frequency statistics shows that weekdays “Tuesday” and “Wednesday” are less frequent than other weekdays. It would be irrelevant, though, to learn frequent weekdays in the beginning leaving the two “infrequent” weekdays for later training. As O’Keeffe et.al. (2007) put it, “pedagogical decisions may override these awkward but fascinating statistics” (p.41).

Nevertheless, in spite of all imperfections of the equation: ‘most frequent’ = ‘most important to learn’ (Leech, 1997, p.16), it is difficult to deny the value of the frequency statistics for selection of leaning materials. It certainly helps separate wheat from the chaff – rare examples and words should be left out for later training (McEnery & Wilson 2001).

There are a number of different corpora for Swedish, among them:

Parole, SUC (two general-language corpora, annotated, written language)

Konkordanser , ORDAT, SNP, Bellman, Strindberg, Litteraturbanken, Press Text, mediaArkivet, eBooklagret, Project Runeberg, FASS, etc. (domain-specific, non-annotated, written language)

Talbanken, ASU, Göteborgs Spoken Language Corpus, etc. (written/spoken production of native speakers and learners, annotated)

CrossCheck, SVANTE, TISUS (written production of learners of Swedish, annotated)

OrdiL – (coursebook texts in Science, Maths and Arts from Swedish compulsory school, non-annotated, domain-specific)

As can be seen, there are only two annotated general-language corpora available for Swedish – Parole and SUC. 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 the size.

A new, modern, large-sized general corpus of Swedish has long been asked for. The initiative has been taken to investigate the need and the possible structure of a potential Swedish National Corpus (SNK). The results of the study are published in Andréasson, Borin, Merkel (2008). Unfortunately, the construction of SNK has not yet received funding and still remains on a wish-list.

To settle the problem of a big modern corpus of Swedish, a web-corpus SweWAC (Swedish Web-Acquired Corpus) has been collected by the KELLY partner "Lexical Computing Ltd" using Corpus Factory tool (Kilgariff, Reddy, Pomikálek, 2010) and is at present available via commercial concordance tool SketchEngine (http://www.sketchengine.co.uk/) as well as a "citation corpus" via . http://språkbanken.gu.se/korp/. The method of collecting a web-based corpus for Swedish consists of several steps:

1. Collect "seed word" list, approx. 500 mid-frequency words whose frequency range is between 1000 and 6000. This is done using texts on Wikipedia – first a "Wiki-corpus" is collected as a primary corpus for seed-word selection, word form frequency is calculated (as opposed to base forms/lemmas), and then 500 mid-frequency word forms are selected for further web-search. Length restriction is 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 are sorted away.
2. Repeatedly select three random seed words to create a query, send query to a search engine.
3. Retrieve hit pages and clean the text, e.g. remove navigation bars, ads, duplicates; check them for the most frequent function words – if they are present, then the page is in the target language. Otherwise, the page is discarded.
4. Tokenize, lemmatize, POS-tag, where possible.
5. Load into a concordance tool.

Web-corpus construction has taken up approximately 2 months and as a result a corpus of 114 mln. words has been provided to the Swedish Language Bank for use in KELLY project.

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

- It is a highly automated process which therefore ensures short collection time at low cost.
- Since the corpus is web-based it is an open-source resource, i.e. presents no obvious copyright problems.
- Texts collected from the web tend to present more spoken-like (interactional) language since there are a lot of forums and blogs and thus,

compared to classical corpora, it has a benefit of complementing strictly written mode of language with everyday-like colloquial language.

Among the disadvantages 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, the text mass is so extensive that there are more chances that there is no skew in favour of any specific topic.
- As our experience of SweWAC has shown, besides texts in Swedish there are 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, Arthur, Berlin, Chris, Chicago, Christian, Charles, David, Daniel; non-Swedish spelling of words, e.g. America (as opposed to the Swedish “Amerika”), British (as opposed to the Swedish “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 5 letters, e.g. album, attack, civil. One way out of this is POS-tagging of the wiki-corpus and filtering seed words of unwanted word classes prior to sending queries to the search engines. Another – even better – alternative is to have a language team prepare seed words for different genres and thus ensure the more or less balanced and predictable structure of the corpus.
- Yet another problem with web corpus is that texts, automatically collected from the web, come in different encodings and it is time-consuming to convert the encoding manually before POS-tagging and lemmatization can be done.

It should be mentioned here that the method of working on the KELLY-lists is formed in such a way that a number of problems mentioned above have been corrected through wordlist comparisons between languages during the post-translation phase. This and some other selection strategies are described later in the text.

2.2 Working with SweWAC

SweWaC has been handed to us after it has been collected with the instrument Corpus factory (Kilgariff et al., 2010). A number of filtering was needed: the encoding of the separate files was different, there were texts in other languages than Swedish (e.g. Danish, Norwegian, English). As long as it was possible, foreign texts have been removed and different encodings converted into. It took about a week of full-time job to make this corpus usable.

2.2.1 Lemmatizing and POS-tagging SweWAC

The raw texts collected through Corpus factory needed to be further lemmatized and POS-tagged before they could be loaded into SketchEngine (Kilgariff et al. 2004). The input format for SketchEngine is one word per line, with tabbed tags and tabbed lemmas, e.g.:

Running word	**POS tag**	**lemma**
Förändringar	NCUPN@IS	förändring

i	SPS	i
departementen	NCNPN@DS	department
:	FI	:
electronic	XF	electronic
edition	NCUSN@IS	edition

Tokenization, pos-tagging and lemmatization were performed using tools developed at Gothenburg University (Kokkinakis & Johansson Kokkinakis, 1997). SweWaC was adjusted to the SketchEngine input format and uploaded into SketchEngine.

At that stage it became possible to take a quick look at the most frequent nouns, verbs and adjectives in SweWAC. The first 25 most frequent nouns, verb sand adjectives suggest that the majority of texts come from newspapers, among them:

- Nouns: år (=year), tid (=time), land (=country), kommentar (=comment), värld (=world), problem, artikel (=article)
- Verbs: säga (=to say), skriva (=to write), tycka (=to consider), veta (=to know), tro (=to believe)
- Adjectives: politisk (=political), ekonomisk (=economic), viktig (=important), svensk (=Swedish), senaste (=recent)

Analysis of hapax legomena as well as 50 longest words in each subcorpus suggests a relatively large amount of Internet-related text types, i.e. forums, blogs, chats, and other online communication. An example is the longest *http-free* word is:

HAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAH
AHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHA
HAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAH
AHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHA
HAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAH
AHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHA
HAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAH
AHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHA
HAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAH
AHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHA
HAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAH
AHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHAHA

Some typical hapaxes are: bravoo, soooooooooooooooooooooooooooo, ooooooooooosnaaaaaaaaaaaaaaaaaaaaaaaaaa (=a variant of ”donkey”?)

Analysis of the first 9000 words performed during the work on the Kelly list supported our initial hypothesis and revealed a mixture of political terminology, historical words as well as everyday expressions.

2.2.2 The notion of “lemma” in the Swedish KELLY-list

Here it is important to comment on what we understand by lemma in this context.

The way researchers operationalize the construct “word” influences the way word statistics and frequency counts are collected and the way different aspects of individual words are analyzed. This has a direct impact upon the pedagogical application of the collected statistics (Gardner 2007). As has been mentioned above, the frequency count in the Swedish KELLY-list is calculated upon lemmas (or lemgrams as they are otherwise called). Lemma is a useful concept for applied corpus studies, but it contains a number of drawbacks. There exist different ways to define the notion of lemma. The way lemmatization has been made in SweWAC (and consequently the way it has been inherited by the KELLY-list) does not exactly reflect the way we would like to define it.

In SweWAC context lemma (lemgram) is understood as 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, flickornas* are counted together since they have the same base form *flicka* (Eng. girl), the same word class *noun* and the same gender *uter*. This is reasonable. 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.

Furthermore, with the exception of a number of very frequent multiword items, most of them are not identified as units, but are rather split into constituent parts and each part is counted separately. Among the exceptions to this general approach we can name *bland annat* (Eng. *among other things*).

Another aspect that is missing in SweWAC annotation 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, v* and *lärare, n* would make the same entry). The frequency statistics collected 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*).

However, errors in frequency calculations 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 aren’t many (Gardner 2007). Thus the two features - having less frequent multiword units 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 lemma both reliable and appropriate for language learning purposes.

2.2.3 Lemgrams: SketchEngine and frequency measures

To create a wordlist SketcheEngine has been used, tab “Wordlist”. Obviously, there is no default option in SketchEngine for creating wordlists containing multiple information, e.g. lemma-tagset-frequency (three parameters at once). The tagset had to be specifically adjusted for various wordlist options by SketchEngine system engineers to make possible creation of wordlists based on the three above-mentioned parameters.

As a result, two wordlists have been generated: one with lemma-tags in combination with raw frequency; and one with lemma-tags in combinations with ARF-frequency (average reduced frequency). ARF takes into account dispersion of the words in different subcorpora and throughout the whole corpus. If the word/lemgram 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 Savichý and Hlavácová (2002).

We generated a lemma-tag list consisting of lemma, tag, and frequency in the following format:

i:-:SPS	1353224.0
en:-:DI@US@S	999690.4
vara:-:V@IPAS	976049.4
det:-:PF@NS0@S	958912.6
som:-:PH@000@S	889632.2

Lemma-tag list with raw frequency has provided us with 402 446 lemmas

Lemma-tag list with ARF-frequency has yielded 232 900 lemmas

Reduction in number of lemmas that qualify themselves for inclusion into basic vocabulary is obvious due to the use of dispersion adjustment.

The first step was to merge raw frequency lists with ARF lists. Raw frequency gives the relative frequency per million words (wpm), which is a comparable value between different corpora. However, since we intended to order the list according to the ARF-frequency, we retained both ARF-score and collected the raw frequency for the items in the ARF-list. The merged list for SweWAC contained 153 061 lemmas.

Once the ARF-list with the relative frequency (raw frequency per million words) has been created, the next stage has started.

2.3 Processing M1 word list

2.3.1 Principles for POS-selection

The main guideline in selecting word classes for the Swedish KELLY-list was a document produced by a KELLY partner “Proposal for inclusion of word types in Kelly” (available on the project wiki-page). According to that document the following should be included:

- base forms with normalized spelling (i.e. lemmas as we understand them, see 2.2)
- no affixes
- derivational forms are legitimate independent items and should not be grouped according to the root morpheme

- abbreviations if they stand for the type of words we include (e.g. no abbreviations for proper nouns)
- multi-word units are not included in the Swedish list except a number of those that are automatically identified; yet if some vocabulary item is ungrammatical when used outside of a phrase, add the context (e.g. “bege” (Eng. “to go”) is not used without reflexive “sig”)
- no idioms or other phraseological units
- no proper names with the exception of geographical names that have gained their place according to the frequency range. Yet, do not include the ones that are typical to the country where the language is spoken; exceptions are the name of the country, name of the people, language, and main cities;

The following word classes were suggested for inclusion: noun, verb, adjective, adverb, pronoun, determiner, conjunction (and subjunction), exclamation, 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).

Exclude: participle, proper nouns, foreign words (if these are annotated as such), punctuation

2.3.2 Identifying and filtering “noise”

30% of 153 061-long list was constituted of “noise” that we removed automatically. By noise we understood the following groups:

a. All entries (lemmas) containing comma (,), full stop (.), semicolon (;), colon (:), asterisk (*), quotation marks (“) and (”), apostrophe (‘), dash/hyphen (-), &-character (&), slash (/ and \), greater-than (>) and less-than (<). We preserved items containing underscore (_) since underscores are used in multiword items (e.g. d_v_s, i_alla_fall). There are some “good” items that have been sorted in the process, for example some abbreviations containing full stops. Yet, the percentage of “rubbish” compared to the “good items” is so high that it was worth doing it.

b. Some word classes:

- Proper names – we have assumed that these were not as important for the learner 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, Sweden), and the main Swedish cities (Stockholm and Gothenburg). Automatic sorting was possible since our tagger makes distinction between nouns and proper names. Had that not been possible, we wouldn’t have made this filtering.
- Numerals have been removed from the list on the assumption that the number of numerals in the list is too high whereas the most necessary numerals (43 of them) could be added manually faster than the rubbish-numbers can be removed manually. Among the added numerals are ordinal numbers 1 to 20, 30, 40…100, 1000, 1000000 plus some cardinal numbers “first”, “second”, etc.
- Punctuation marks have been removed.
- Participles have been 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 Academic Word List), do not provide participles as separate entries; they are, instead, listed together with the verb.

- Foreign words that have been recognized by the tagger, have also been removed.

Altogether 51 522 lemmas have been removed as ”noise” reducing the original 153 061-long list to approx 100 000-long list of lemmas.

Final reduction in lemma-number was done automatically by collecting all morphological variants of the same lemma under one unique entry. As an example, the original list contained all forms of the adjective “livlig” (=”lively”):

lemma:-:POStag	ARF	RF	Word form
livlig:-:AQPUSNIS	270.9	RF=450	livligt (neutrum)
livlig:-:AQP0PN0S	168.2	RF=284	livliga (plural)
livlig:-:AQPNSNIS	60.3	RF=94	livlig (utrum)
livlig:-:AQC00N0S	53.1	RF=77	livligare (comparative degree)
livlig:-:AQS00NDS	19.5	RF=2	livligaste (superlative degree)

All forms identified as “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	{tags={arf=rf}}
572	907.0	7.955	livlig	AQ	{AQC00N0S={53=77}, AQP0PN0S={168=284}, AQPNSNIS={60=94}, AQPUSNIS={271=450}, AQS00NDS={20=2}, subtotal={572=907}}

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

To go through a list of 54 000 lemmas isn’t an easy task, therefore we cut the list at 9000-point and started with this. The reason for having 9000 cutoff-line is that the final list for language learners should be 9000-lemmas long, even though the first list for translation should be 6000-long. However, in case the translation would not be able to enrich the original list with the rest 3000, we will have some extra items to collect from a cleaned and proofread list.

This list containing 9000 items was the one we started working with.

2.3.3 Abbreviations

We have decided to follow the following “mode” of presenting abbreviations:

bland annat (förk. bl.a.)	adverb
BNP (bruttonationalprodukt)	noun-en
kilogram (el. kilo; förk. kg)	noun-ett
kl. (klockan)	noun-en
sankt (förk. s:t)	adjective

Table 1. Examples of abbreviations

In most cases it is the full form (the way the abbreviation is pronounced or read aloud) that is used as a headword, and in brackets the abbreviation or several variants of the abbreviation are provided, see “bland annat” (Eng. “among other things”). The word “förk.” stands for “förkortning” (Eng. “abbreviation”).

Another case is when the word is abbreviated but it is normally pronounced as the letters constituting the abbreviation, e.g. “BNP” is pronounced /be en pe/. In such cases we have used the abbreviation as a headword and provided the full word in brackets, see “BNP” (Eng. “GDP”, “Gross Domestic Product”) in the table 1 above.

There are cases of the type “kilogram”. Kilogram can be shortened to “kg”; another way of writing it is “kilo”. All the three variants are used in the corpus. Difference between “kg” and “kilo” is that “kg” is never pronounced as /ke ge/ – it is extended to its parent form “kilogram”. “Kilo”, on the other hand, is pronounced as /kilo/. Therefore marking both “kilo” and “kg” as abbreviations is not consistent. The learner might make a conclusion that the form “kilo” is pronounced “kilogram”. Or vice versa, that “kg” should be pronounced as /ke ge/. Splitting the entry into two – “kilogram (el.kilo), noun” and “kg (=kilogram), förk.” is wasting a valuable entry (since we are allowed to keep only 9000 entries in our final list. The alternative we followed is to have an entry containing all information “kilogram (el. kilo; fork. kg), noun”.

Yet another abbreviation case can be exemplified by the case of “kl.” (Eng. o’clock). The problem with this item is that if we use the full word “klockan” as the headword, it will go against the lemma-rule. “Klockan” is a definite form of the word “klocka”. We cannot use “klocka” (Eng. 1. clock; 2. bell) as the headword since it is not abbreviated to “kl.” in all meanings. “Kl.” can only be used with reference to a definite point of time (e.g. “kl.17.00”). We have therefore kept the form “kl.” as the headword with its parent form in brackets to avoid misleading interpretations.

2.3.4 Proper names

The decision to filter all proper names has been dictated by the fact that most proper names among the first 10 000 words are person names that are not of much interest for the learner. There have been arguments that the first 10 male names and the first 10 female names may be useful to know for someone who studies the language. At the same time the primary application of the list is for flash cards, which means every item should be matched with its translation. Pairing person names with their translations does not sound as a relevant task, though.

The rest of the proper nouns have been filtered on the assumption that city names and country names for the partner languages/countries can be added manually or can come into the Swedish list at the stage when we start merging the master list with the translations from other languages into Swedish. It is faster to work this way than to delete numerous proper names manually from a 9000-long wordlist.

The lemmatizer has missed to mark certain proper names correctly, and we received a list containing for instance lemmas “skatteverk, noun” (Eng. “a tax department”) and “migrationsverk, noun” (Eng. “a migration office”). The correct proper names should be “Skatteverket” and “Migrationsverket”. After the discussions whether these were of sufficient value for the learner to keep in the list, we have decided that proper names denoting the social structure of a country are domain-specific and cannot be called “base vocabulary”. If on the next stage we decide to include words of this domain into KELLY-lists, we will identify the necessary words and add them manually.

2.3.5 Spelling and form variants. Introducing “lexicographic” approach

Working on the assumption that this list is more of a descriptive character rather than prescriptive (presumably that is why we are working with corpora rather than our intuitions), we have not taken away different spellings or different forms.

The original idea was to use the most frequent form or spelling variant as the headword providing other variants in brackets. In many cases different variants (including spelling variants) of the same words gained their own entry in the list before we started proofreading it, e.g. “far” and “fader” – two variants of the word “father”. We have merged the two entries first following the principle of the most frequent variant being given the status of a headword and providing the second variant in brackets.

However, we could not follow the principle of “most frequent makes the headword” consistently. The main reason for this was that entries would be inconsistent in case of several parallel cases. For example, in the case of “far” and “fader” (Eng. “father”) the most frequent is “fader”, while for “mor” and “moder” (Eng. “mother”) the most frequent is “mor”. Since the two cases are obviously parallel in nature, to use “fader” as a headword in one case and “mor” in the other, does not rend consequence to our list. At least, we felt that this will be confusing to the end-user. We had in the end to abstract from the statistics and go for the lexicographic principle using a more neutral alternative as a headword in all cases.

There are parallel cases of the same type with words for “grandfather” (morfar/morfader and farfar/farfader), “grandmother” (mormor/mormoder and farmor/farmoder), “uncle” (farbror/farbroder and morbror/morbroder), “brother” (bror/broder). To keep some consistency in the list we used the short form as the main form and the longer form as the second alternative.

This has rendered us with the entries as shown in Table 2:

en	mor (el. moder, vardagl. morsa)	noun-en	mother
en	far (el. fader, vardagl. farsa)	noun-en	father
en	bror (el. broder, vardagl. brorsa)	noun-en	brother
en	syster (vardagl. syrra)	noun-en	sister
en	farbror	noun-en	uncle
en	morbror	noun-en	uncle
en	morfar	noun-en	grandfather
en	farfar	noun-en	grandfather
en	mormor	noun-en	grandmother

en	farmor	noun-en	grandmother

Table 2. Examples of parallel cases having alternative form variants

In many other cases where more than one spelling variant was present in the frequency list it was quite straightforward. We went in the first hand after the spelling provided in SAOL (Swedish Academic Word List) that has been used as our primary reference source. The less prevalent variant according to SAOL was provided as an alternative spelling. Some examples are given in table 3.

buddhism (el. buddism)	noun-en	buddism
jävla (el. djävla)	adjective	bloody
karaktärisera (el. karakterisera)	verb	characterize
kilogram (el. kilo; förk. kg)	noun-ett	kilogram
klä (el. kläda)	verb	to clothe, to dress
ner (el. ned)	adverb, particle	down
numera (el. numer)	adverb	nowadays
så småningom (el. småningom)	adverb	eventually
ta (el. taga)	verb	take
television (el. teve, tv)	noun-en	television
timme (el. timma)	noun-en	hour

Table 3. Examples of alternative spellings

One more group with spelling variants is a large group of multiword expressions that can be spelt as several words or as one word in Swedish. Here you can find “first of all” (framförallt/framför allt), “sometimes” (ibland/i bland) and a number of other similar cases. We could not consult SAOL in these cases since it does not contain multiword expressions. In these cases we followed the principle “most frequent merits the headword status”. Some examples of those follow below in table 4.

allt mer (el. alltmer)	adverb	increasingly
framför allt (el. framförallt)	adverb	above all
hur som helst (el. hursomhelst)	adverb	anyway
i alla fall (el. iallafall; förk. iaf)	adverb	in any case
i fråga (el. ifråga)	adverb	in the question of
i gång (el. igång)	adverb	running
i morse (el. imorse)	adverb	this morning
ibland (el. i bland)	adverb	sometimes
igår (el. i går)	adverb	yesterday
ikväll (el. i kväll)	adverb	tonight
istället för (el. i stället för)	prep	instead of
tvärtom (el. tvärt om)	adverb	on the contraty

Table 4. Examples of multiword expressions with alternative spelling

2.3.6 Homonymy, polysemy

Some teams within the project have decided to disambiguate homonymous and in certain cases polysemous items prior to the translation phase to avoid multiple

translations of the same entry. The Swedish team has decided to go after the lemma-principle to make the process more automatic and fast. The lack of time we experienced was mainly due to the fact that we did not have the core corpus available from the beginning and it was unclear how fast it will be delivered to us. On the other hand, it was also a part of the decision to run an experiment that will help identify how many one-to-one mappings there are between different languages; how homonymous and polysemous items can expand after the translation; and how much in percent the list will 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 didn’t intend to limit the translations by the provided context (table 5).

	ens	adverb	e.g. inte ens ngt/ngn, med ens	even; at once
en	fan	noun-en	e.g. sportfans	fan
att	gifta	verb	e.g. gifta sig, gifta bort	marry
att	haka	verb	e.g. haka av/fast/på	to hook
att	hamna	verb	e.g. hamna i/på	to land
	ju	conj	e.g. ju mer…desto bättre	the (more, the merrier)
	medelålder	noun-en	e.g. medelåldern, medelålders	middle age
	si	adverb	e.g. si och så, si så där (el. sisådär)	so
	vis	noun-ett	e.g. på så vis/på sätt och vis	way
	övrigt	adverb	e.g. i övrigt, för övrigt	otherwise

Table 5. Examples of items followed by “example” column

We left a lot of disambiguation decisions to the translators. One example of those is the headword “rom”. In different contexts it can mean a drink (Eng. rum), caviar, a collective name for gypsy people, or a city (Rome). In all the cases the noun is used without articles, and is of a non-neuter gender (takes article “en”). The rule of the thumb for translators has been to use the most frequent alternative and to keep in mind that the list is intended for language learners.

In the first version of translations the following interpretations have been provided for Swedish “rom, n-en”:

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

Table 6. Translations of the Swedish item “rom” into the 6 European Kelly languages

According to the provided translations, the equivalents for the Swedish “rom” in the partner languages are mostly used as a drink, caviar or roe deer; none of the

translators has offered the alternative for the name of the city (probably because of the word class. City names should be marked as proper nouns), nor the collective name for gypsies. The translator into Russian has shown a good sense of humour choosing alcoholic drink as the most relevant sense for language learners. The translation paradigm shows that the translated items cannot be used as translations of each other.

In table 7 below we have collected some information on multiple translations (several translation equivalents for one and the same item) in different languages. There are even a number of comments from translators that often explain why certain items haven't been translated into the target language.

Language	Cells with multiple translations (homonyms)	Cells with comments
English	319	20
Greek	1021	493
Italian	857	21
Norwegian	1	0
Polish	325	32
Russian	7	52

Table 7. Number of multiple translations from Swedish into the 6 European Kelly languages

2.3.7 Stylistically marked versus neutral vocabulary

In the guidelines we have defined that the basis for the KELLY-lists should be neutral vocabulary. It is, however, very difficult to neglect the frequency statistics. Therefore the Swedish list covers a number of entries that contain stylistically marked words. They are of two kinds.

1. The initial unprocessed Swedish list contained a lot of "duplicate" entries: words that are in fact variants of each other have gained individual entries due to lemmatization, e.g. "dem" (Eng. them) and "dom" (colloquial variant of "dem"). We merged these manually where we could discover repetitions of this kind. In this case first comes the neutral item (headword) followed by the stylistically marked variant in brackets. The non-neutral variant is then preceded by one of the markers – "vardagl." (Eng. colloquial, everyday-like) or "formellt:" (Eng. formal). Some examples are shown in table 8:

	allihop (vardagl. allihopa)	pronoun	all; everyone
	alltihop (vardagl. alltihopa)	pronoun	all
	de (vardagl. dom)	det	the
	de (vardagl. dom)	pronoun	they
	dem (vardagl. dom)	pronoun	them
	dig (vardagl. dej)	pronoun	you
att	fungera (vardagl. funka)	verb	work
att	ge (formellt giva)	verb	to give
	inte (formellt: icke, ej)	adverb	not
	medan (vardagl. medans)	subj	while

	mig (vardagl. mej)	pronoun	me
	någon (vardagl. nån, förk. ngn)	pronoun	someone
	någonsin (vardagl. nånsin)	adverb	ever
	någonstans (vardagl. nånstans)	adverb	somewhere
en	socialdemokrat (vardagl. sosse)	noun-en	Social Democrat
en	syster (vardagl. syrra)	noun-en	sister (informal: sis)

Table 8. Examples of items with stylistically marked variants

2. Another category of stylistically marked words is presented by the group of words where the headword itself is non-neutral. In this case we marked that in a separate column. The following range of stylistic markers is present: “vardagligt” (Eng. colloquial), “stötande” (Eng. offensive), “ålderdomligt” (Eng. archaic, old-fashioned). Some examples of such words are provided below in table 9:

en	grej	noun-en	(vardagligt)	thing
en	hora	noun-en	(stötande)	whore
	info	noun-en	(vardagligt)	info
en	jävel	noun-en	(stötande)	bastard
	jävla (el. djävla)	adjective	(stötande)	bloody
en	koll	noun-en	(vardagligt)	check
att	kolla	verb	(vardagligt)	to check
att	käka	verb	(vardagligt)	to nosh
	less	adjective	(vardagligt)	sick and tired
en	skit	noun-en	(stötande)	shit
att	skita	verb	(vardagligt)	to shit

Table 9. Examples of items with stylistically marked items

3. The last group of stylistically marked words is constituted by a small group of interjections that are highly colloquial. They have been deleted manually during the initial processing of the list in accordance with the Athens agreement on word inclusion; moreover, these interjections are very specific for Swedish, they do not have much learner value and it is not clear how some of them can be translated, see some examples of those follow in table 10:

nja	interj	well,ok… (reluctant acceptance)
oh	interj	oh
jodå	interj	yeah
hm	interj	hm
jaja	interj	well, well
å	interj	and; to (inf marker)
hmm	interj	hmm
ah	interj	ah
eh	interj	eh
jaså	interj	oh; really; I see

wow	interj	wow
äh	interj	errr…
sååå	interj	sooooo

Table 10. Examples of colloquial items that have been removed from the Swedish Kelly list

2.3.8 Multiword expressions

The multiword expressions are a special case for automatic identification and tagging. In Athens we agreed that certain language teams will take care of those, but not every team. The Swedish team decided to accept those multiword expressions that can be automatically identified. There were 154 such items in the 6000-long list that was sent to translators. A number of other items that could not be automatically identified but were manually discovered during the proofreading stage, have been fixed in the list. The most numerous group here consisted of reflexive verbs (16 of them), e.g. “nöja sig” (Eng. enjoy oneself). We haven’t used the term “multiword expression” as a word class. Instead, these items are classified either as adverbs, conjunctions, prepositions, pronouns or verbs, see some examples in table 11.

så gott som	adverb	as good as
på grund av (förk. p.g.a, pga., p g a)	adverb	due to
på något vis	adverb	somehow
söder om	prep	south of
tack vare	prep	thanks to
trots att	subj	in spite of
var och en	pronoun	each and every
vare sig	conj	neither
bete sig	verb	to behave
bosätta sig	verb	to take up residence
bry sig	verb	to care
förhålla sig	verb	to be related; to take a position

Table 11. Examples of MWE of different word classes

2.3.9 Borderline cases

A number of decisions had to be taken as far as treatment of borderline cases was concerned. Some cases are exemplified here:

1. We decided against keeping gender distinctions, e.g. “svensk (male, noun)” versus “svenska (female, noun)”.
2. Some forms of pronouns have been given individual “headword” status instead of being reduced to the same lemma, e.g. “alla” (Eng. everyone, plural), “all” (Eng. everything, non-neuter gender, singular), “allt” (Eng. everything, neuter gender, singular). We have manually merged the three of those, summing up their frequencies into the entry

all	pronoun

3. The question of adverbs derived from adjectives with the suffix “t” has been discussed as to whether these should be merged into one entry or kept apart. The following arguments have been raised. On the one hand, it can be argued that adverb-building suffix “t” is a derivational suffix. Then if we follow the principle of merging derivational forms into one entry, other “simple” derivational patterns should be abandoned as well, e.g. derivational prefix “o-“ that builds negative form of a word, e.g. “bekväm-obekväm” (Eng. comfortable-uncomfortable), adjective-building suffixes “-full” and “-lös”, e.g. “meningsfull-meningslös” (Eng. meaningful-meaningless) and the like. How should we then select between “simple” patterns and more “complex” ones? Should we abandon all of the derivational patterns? What about compounding? If we keep to that principle, we will sooner or later reduce our list to a list of “word families” united by the same stem, which goes against the principles agreed upon by the partners in Athens.

   On the other hand, in Svenska Akademiens Grammatik (Eng. ”Swedish Academic Grammar”, SAG) “traditional” adverbs that have been derived from adjectives by adding suffix “t” are no longer counted as adverbs. They are nowadays treated in SAG as a neuter form of adjectives used in adverbial function in the sentence. Nothing is said about the adjectives and adverbs that have the same graphic form, though (e.g. exakt, konsekvent, korrekt). However, in the Swedish Academy Word List such “adjective-adverb duplicates” have alternately one or two entries (either adjective only; both parts of speech in the same entry or as two separate entries for different parts of speech). It is therefore difficult to generalize. Still, if we follow SAG principle in viewing “-t” derived adverbs not as adverbs, but as a specific form of an adjective used adverbially, we can distance ourselves from the “word-family” trap and additionally get 157 empty entries for extra vocabulary. It was deemed reasonable since the core meaning in adjectives without “-t” is preserved when “-t” is added.

   However, after consulting the representatives of other partner languages we decided to keep adjectives and “t”-derived adverbs as two separate entries since in the other languages these pairs can yield different translations.

   Some examples of adjective-adverb pairs:

Adjective-adverb	Translation
Aktiv – aktivt	Active – actively
Allvarlig – allvarligt	Serious – seriously
Evig – evigt	Eternal – eternally
Grov – grovt	Coarse – coarsely
Låg – lågt	Low - low

Table12. Examples of t-derived adverbs

2.3.10 Proofreading

MANUAL PROOFREADING 1

Lemmas and their word classes were checked manually word by word. SketchEngine was regularly used to see the examples of actual use of lemmas in concordance forms. SAOL was consulted regularly for morpho-syntactic information, spelling, and in certain cases for existence of this or that lemma in the dictionary.

A lot of cases of varying spelling variants have been reduced to one entry, as for example in the case of ”så_kallad”, ”så_kallat”, ”så_kallade” and ”sk” (Eng. ”so called”), se even 2.3.5. Frequencies for all these variants have been summed up and the four entries reduced to one lemma ”så_kallad”.

Errors made by the tagger have often been identified, e.g. in case of the noun “spök”, which has been corrected for “spöke” (Eng. ghost). Some cases were more difficult to identify, e.g. in case of “fångare”. SAOL doesn’t contain “fångare” and we therefore had to make a decision whether to delete the word altogether or identify the form that actually exists. A check in the SketchEngine concordance revealed that “fångare” was a mistakenly identified lemma for the plural form from “fångarna”. Thus, this entry was changed for “fånge” (Eng. prisoner) and later had to be merged with an already existing “fånge”, adding the frequency to the total for this lemma.

The list was somewhat shortened in this (manual) way leaving after the first proofreading 8455 lemmas.

CHECKING AGAINST OTHER DICTIONARIES (SALDO, SAOL)

It is easy, however, to make omissions during a manual control. Therefore, to double-check that the resulting list contained only existing words, a matching against an existing online dictionary SALDO (Borin et al, 2008) freely available from the Swedish Language Bank was performed. About 500 warnings were issued. The list of those items were double-checked manually – certain passive verbs that didn’t contain suffix “s” were corrected, e.g. “envisa”→ “envisas” (Eng. to persist); some reflexive verbs have been corrected for “sig”, e.g. “befinna” → “befinna sig” (Eng. to be present), some forms have proven to be existing via an extra check in SAOL; others seemed to be very modern to be included even in SAOL, e.g. “blogginlägg” (Eng. blog entry).

MANUAL PROOFREADING 2

The last check that was performed before sending the list for translation was a human check where apart from controlling different inconsistencies, articles had to be assigned to nouns (in case they are used with articles), and infinitive markers to be set prior to verbs (if applicable). This last control raised a number of issues that led to further improvements of the list.

2.3.11 Adding items manually

88 items were manually added, among those 43 numerals and words for

- body parts: *ögonlock*.
- countries/geographic names (for all 9 members of this project): *Europe*, *Greece*, *Great Britain*, *Italy*, *China*, *Norway*, *Poland*, *Russia, Gothenburg*, *Stockholm*, *Sweden*;
- directions: *bakre*, *främre*, *nordost*, *syd*, *sydväst*, *sydöst*.
- family members: Most family members were already on the list, the only ones we added were *morbror*, *storasyster*, *syssling*, *änkeman*, *änkling*;
- Among holidays only *pingst* (=Trinity) was added; *Jul* (Christmas), *nyår* (=New Year), *påsk* (=Easter) and *midsommar* (=Midsummer) were already in the list through the original list based on the corpus.
- Meals: *brunch*, *kvällsmål* and *mellanmål* were added;

- Measure words – *decimeter*, *hekto*, *hela*, *milligram*, *millimeter* were added;
- Politics: *inrikesminister* (=Minister of Domestic Affairs), *inrikespolitik*, *utbildningsminister*, *vänsterpartistisk* were added;
- days of the week: *fredag* had to be added manually – the rest of the weekdays already entered the list through the frequency statistics from the Swedish WaC.
- senses: *hörsel* and *känsel* were added;
- time: *månadsskifte* and *veckodag* were manually added.

If on some stage we should decide not to keep the words that were manually added, they would have been easily removable since they are marked as "manually added" items in the "Source"-column.

The manually added items were assigned the highest frequency to make sure that these words come first since they have high learner importance.

2.4 Raised problems

2.4.1 POS between languages

It was agreed that we would not ask translators to mark which part-of-speech the translated item belongs to. Presumably, the source item POS can be assigned to the translation items. However, this approach has some pitfalls. A short check of the source lists from other languages has shown that the taxonomy of word classes differs in different languages, see table 13 below:

Arabic	Chinese	English	Greek	Italian	Norwegian	Polish	Russian	Swedish
No POS	No POS	Abbr.	Abbr.	Abbr.	-	Abbr	-	Abbr.
		Adj.	Adj.	Adj.	Adj.	Adj	A	Adj.
		Adv.	Adv.	Adv.	Adv.	Adv	R	Adv.
		Conj.	Conj.	Conj.	Conj.	Conj	-	Conj/subj
		Determ.	Article (2)	Determ.	Determ.		-	Determ.
		Exclam.	Exclam.	Interj.	Interj.	Interj	-	Interj.
		-	Expression	-	-	-	MWE	-
		Miscel-laneous	-	-	-	-	-	-
		Modal verb	**-**	**- for**	-	**- vr**	-	Aux. verb
		Noun	Noun	Noun	Noun	Noun	N	Noun-en, noun-ett
		Number	Numeral	Numeral	-	Num	M	Num
		-	Particle	-	-	Particl	Q	Particle

						e		
		Prep.	Prep.	Prep.	Prep.	Prep.	-	Prep.
		Pron.	Pron.	Pron.	Pron.	Pron.	-	Pron.
		-	-	Proper noun	-	-	-	Proper noun
		Verb	Verb (2 diff types)	Verb	Verb	Verb	V	Verb

Table 13. POS taxonomy across KELLY languages

Arabic and Chinese didn't assign any word classes at all. Certain word classes in other languages did not have correspondences in Swedish, e.g. "MWE" or "Miscellaneous" in Russian and Greek. Numerals in Norwegian were listed among determiners. Practically viewed, it meant that when the check between the lists started, it was not obvious whether the lemgrams from original Swedish list and a translation from another language into Swedish would match each other and give the necessary "score" for the item, e.g. Swedish "arton, numeral" (Eng. eighteen) and translation from Norwegian "arton, determiner". The decision in the end was to drop POS and compare lemmas only instead of lemgrams, as will be described in Chapter 3.

2.4.2 Prescriptive versus descriptive list

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 cards for the 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 aim to use a modern corpus of Swedish to identify lexical items that are frequent in modern Swedish and which therefore are necessary for the language learner to study in the first hand. Thus, if we started applying "selection" rules based on our judgment rather than statistics, it would be a step back and we risked ending up with a regular list.

On the basis of this, we decided not to delete certain vocabulary that didn't look "appropriate" for language learners, e.g. words like "stalinistisk, adj", "marxistisk, adj", "sovjetisk, adj" and the like. During the "post-translation" stage we intended to evaluate every dubious item in the list against translations into Swedish, and if any of the above-mentioned words had been used in all (or most) of the lists, they would be assigned a certain score and get an "upgrade" to basic-vocabulary status and therefore would be kept in the final list. If, on the other hand, no other list contained these words, they would be "degraded" and most probably not qualify themselves into the final list at all. Such an approach promised objectivity and consequence in handling ALL items, and not only the ones that seemed out-of-place to us at the initial stage.

2.4.3 Core vocabulary versus domain vocabulary

The problem with core vocabulary falled into two distinct parts:

1. Words that were absent in the Swedish list but should have been included, according to our judgment
2. Words that were present in the list, but which our intuitions say shouldn't have been there

1. After the first draft of the list was done, our attention was drawn to the fact that a lot of words important for language learners were not on the list, among others ‘an orange’, ‘an elbow’, ‘a banana’, ‘an alphabet’. Discussion with other partners has shown that many of them were also concerned with the same problem. One of the reasons for lacking vocabulary might be the fact that internet corpora are not ideal for the language learning purposes when it comes to down-to-earth everyday vocabulary. Another reason might be that such words are too domain-specific and therefore haven’t merited a place in a base vocabulary.

It doesn’t justify, though, the vocabulary absence in the lists. First of all, food, body parts and a number of other topics are specifically named in CEFR descriptors. Second, this vocabulary is often crucial when settling essential needs. That is why a number of partners started comparing wordlists from web-corpora with more learner-oriented corpora. During the post-translation stage some of the lacking words were expected to come into our lists through translations.

The “list-enriching” strategy had been under the discussion for quite some time. Logically, CEFR descriptors should have been the leading guide for preselecting extra vocabulary for inclusion. According to the CEFR document (Council of Europe, 2001) there are four main sources of vocabulary that should/could constitute the vocabulary scope of a CEFR-based course, namely:

(1) words typical for the topics required for the learners’ communication (domain-specific vocabulary);

(2) vocabulary that is based on lexical-statistical principles of selection (highest frequency words);

(3) words randomly coming from texts that are selected as learning material by teachers and finally

(4) words learnt in response to the communicative needs that arise.

It is clear that the second source of words (lexical-statistical) is being taken care of in the KELLY-lists. It is logical then to ensure that a number of topic-specific words that are named as the first source find their way into the list. The last two sources will mainly depend on the teachers that select learner texts and on the communicative situations that the students will be actually engaged in.

Domains and even topics are identified in the CEFR document, even though a bit vaguely. But there are questions about how many words from each domain/topic? Which words to include and where to draw the line? On which principle should they be assigned to the CEFR levels, if the frequency principle cannot be followed?

We in the Swedish team found an escape in the CEFR document itself Council of Europe 2003, p. 52-53):

> Clearly, this particular selection and organisation of themes, sub-themes and specific
> notions is not definitive. It results from the authors’ decisions in the light of their assess-
> ment of the communicative needs of the learners concerned...
> Users of the Framework, including where possible the actual learners concerned, will of
> course make their own decisions based on their assessment of learner needs, motiva-

tions, characteristics and resources in the relevant domain or domains with which they
are concerned For example, vocationally-oriented language learning (VOLL) may develop
themes in the occupational area relevant to the students concerned. Students in upper
secondary education may explore scientific, technological, economic, etc. themes in
some depth. The use of a foreign language as medium of instruction will necessarily
entail a close concern with the thematic content of the subject area taught."

In view of that, our list can be viewed as recommended general-purpose vocabulary for any type of CEFR courses, and domain vocabulary can be added by a tutor based on the aims of each particular language course. List of domain words can be compiled separately and provided as support for language tutors.

It should be named, though, that during the post-translation stage some of the lacking words came into our lists through inclusion candidates.

2. The problem with vocabulary that is included in the first version of the lists but isn't really appropriate for the learner can be demonstrated on the example of the words "bolsjevik" (Eng. bolshevik) and "stalinism" (both were present in the Swedish M2 list). A number of other words that had been pointed out to us as inappropriate were Europaparlament (Eng. European parliament), Europakonvention (Eng. European convention), Europaråd (Eng. European Council), Mellanöster (Eng. Middle East), and political words typical for Swedish, e.g. typical political parties. Almost all of the mentioned words were removed from the list on the basis of that they are proper names that were erroneously tagged as nouns. Since we kept to the strategy to delete proper names from the list, it was a justifiable decision. However, words denoting parties that are no longer proper names but can be used in other languages were kept, as well as some other political words; see some examples of those in table 14 below.

Article	Word for translation	Example	Article	Translation
ett	arbetarparti		a	labour party
ett	centerparti	e.g. Centerpartiet	a	centrist party
ett	folkparti	e.g. Folkpartiet	a	popular party
ett	kommunistparti		a	Communist party
ett	miljöparti	e.g. Miljöpartiet	a	green party
ett	piratparti	e.g. Piratpartiet	a	pirate party
ett	riksdagsparti		a	party in the Riksdag
ett	vänsterparti		a	Leftist party
	stalinism			stalinism
	stalinistisk			stalinist
en	bolsjevik		a	bolshevik
	marxism			marxism
en	marxist		a	marxist
	marxistisk			marxist
	nazism			nazism
en	nazist		a	nazist

	nazistisk			nazist
	rasism			racism
en	rasist		a	racist
	rasistisk			racist

Table 14. Examples of "political" items in the Swedish M2

Obviously, these words are domain-specific. At the same time since they had the frequency range that allowed them into the first version of the list, we had no "legitimate" ground to remove them from the list. Another reasonable question that we asked ourselves before we took a decision was: if the above-mentioned words are not "approved" for the list, what about words like "socialism", "socialistisk", "nationalism", "nationalistisk" etc.? Shoud these also be removed on the basis of being too political? We had chosen to keep to the principle of objectivity: we kept these words on the pre-translation stage, but controlled their presence/absence in the translations from other languages.

3. Post-translation phase.

3.1 Some words on translations

The quality of translations was very varied. Certain lists had to be retranslated, others – proofread, and only a few were accepted in the first version. Even after those extra steps we discovered quite many spelling and lemmatizing mistakes in translations to Swedish, table 15 provides some examples:

FROM language	**Wrong form/spelling**	**Correct form/spelling**	**English variant**	**Nr of discovered errors vs corrected ones**
Arabic	besättnning angår besegrar honom	besättning angå ???	crew to regard win(s) over him	not done
Chinese	anleding det var en dag … hur kan man göra	anledning ??? ???	reason it was a day … how can one do	not done
English	läxor fotografiering mile	läxa fotografering mil	homework photography mile	84/43
Greek	ecyklopedi förburka ordspåk	encyklopedi förbruka ordspråk	encyclopaedia exhaust, drain proverb, saying	105/26
Italian	universal vänsterblocket	universell vänsterblock	universal left block (parties)	101/31
Norwegian	dumheter gammeldags skådespelet	dumhet gammaldags skådespel	stupidity old-fashioned play, spectacle	93/42
Polish	avdeling bassang sammenfattning	avdelning bassäng sammanfattning	department (swimming-)pool summary/abstract	192/122
Russian	arving föhållande omöjligthet	arvinge förhållande omöjlighet	(an) heir relationship impossibility	156/34

Table 15. Examples of translation mistakes

3.2 Kelly Database

To make it possible to store, analyze and compare the nine original lists and their translations a special database - Kelly DB - was created. Two partners - Lexical Computing Ltd and University of Leeds – took that task on themselves, and by

the end of September 2011 the first version was ready for testing. The web-address to Kelly DB is http://kelly.sketchengine.co.uk/. It is open for public use as a look-up resource.

The main reason for the database was to match original lists for each language with the eight translations to these languages to see how many words are represented in all 9 languages (symmetric translations), how many are common to 8 languages, etc. and to generate the following lists:

- Words universal to all 9 languages
- Words unique for each individual language
- Words specific for each individual language pair

One can type in a word of interest and see whether it is present in the database and how it is translated into other languages.

Apart from that, the database facilitated generation of the following lists necessary for post-translation editing of the monolingual lists:

- Candidates for exclusion for each individual language, i.e. words present in the target monolingual list (M2) but not used in any of the translations from other languages to the target language
- Candidates for inclusion, i.e. words that have been used in the translations to the target language, but not present in the original list (M2)
- Multiword expressions not present in the original list (M2), but given as translations into the target language from other languages

3.2.1 POS taxonomy

As mentioned in the previous chapter, two of the master lists (M2) – Chinese and Arabic – did not contain any word class markers. Another problem with POS was that certain word class distinctions were absent in certain lists (e.g. proper names, numerals, reflexive verbs, multiword expressions). Apart from that the names used for different word classes differed from language to language.

To avoid problems with parts of speech correlations, it was decided that only base forms without word class distinctions should be imported to the database. That, of course, has its drawbacks, since homographs from different parts of speech count for one item in the database with links to multiple translations. Yet another problem is that candidate lists for inclusion contained base forms that cover several possible lexical items, e.g. a lemma “fire” might mean both “fire, noun” and “fire, verb”, though it is not certain that both the noun and the verb had been used in the translations. On the other hand that was a way to enrich the master list with more vocabulary.

3.2.2 Normalization and DB rules

The first experiment with matching gave poor results. The reason for that was partly different encodings and partly different formats for lemma entries. For example, Swedish entries had often extra information in brackets that didn’t match with any of the translations not containing brackets. It was agreed that in order to improve matching results, each partner should analyze the original list and the way entries are organized, as well as translations into the target language and suggest a number of rules to apply to the entries before they are imported to DB.

The general rules which applied to all lists:

1. all strings are converted to the lower case (there is a lot of inconsistency in capitalisation)
2. the semicolon is used for separating translation variants
3. commas and forward slashes are converted to a semicolon unless requested in the rules below
4. extra spaces are deleted
5. numbers followed by dots or parentheses are replaced with semicolons (in case senses are numbered, e.g. 1. corner 2. angle)
6. the content in parentheses is deleted (unless requested in the rules below)
7. occasional ! at the end of the entry is deleted

Translations to Swedish were handled in the following way:

1. 'att ', 'en ', 'ett ' at the beginning of a translation string is deleted
2. all exclamation marks and question marks are deleted
3. all square brackets are deleted
4. forward slashes are ok to keep - we don't have any in the original Swedish, but there are MWE in English and Polish that use "/" for separating "smb/smth
5. in Greek-Swedish a number of items start with ";" that should be removed.

The points below referred mostly to translations from Arabic and Chinese (with some exceptions)

6. "och de ", "och den ", "och det ", "och en ", "och " are deleted at the beginning of the translation string unless they stand for the entry itself.
7. "dess " is removed at the beginning of the translation string unless it stands for the whole entry
8. "bli " and "blir " are removed at the beginning of the translation string for better matching, unless it stands for the whole entry
9. "av ", "bara ", "bra ", "denna ", "deras ", "dess ", "din ", "dina ", "hans ", "har ", "hennes ", "liten ", "litet ", "min " in the beginning is to be removed except when they stand for the entry itself
10. "de ", "den " "det " to be removed from the beginning EXCEPT when they are followed by "här", "där", "bästa" or stand for the entry itself
11. "ha " to be removed from the beginning EXCEPT in the expression "ha rätt" or when it stands for the entry itself
12. after all this has been removed - remove "är " from the beginning of the strings

3.2.3 Fixing other problems

As mentioned in 3.1, we discovered quite a number of spelling and form mistakes in the translations. To work more effectively with the problematic items on translation lists we ran an automatic match for all base forms from all translation lists into Swedish and identified those that have been used in one language only; those lemmas were automatically matched against an existing dictionary and the ones that didn’t get a match were marked for manual check. There were multiple

cases of morphologically inflected forms, spelling mistakes, multiword expressions and non-existent words. Where possible, the corrections were introduced.

The corrected lists were imported to the Kelly DB.

3.3 Finalizing master lists: from M2 to M3 lists

The Swedish M2 list contained 6000 items. After processing candidate lists it expanded to 8425 items which roughly confirmed the intuitions that translations from other languages could enrich monolingual lists with 2000-3000 items.

3.3.1 Domain vocabulary

It was agreed that each individual language team will decide how to handle the lacking domain vocabulary. Some teams compiled lists of domain words that they planned to add manually to the M2 lists, others planned to use available dictionaries with domain-marked vocabulary to compile such lists for adding. The Swedish team decided to add the domain vocabulary that came through translations, even though it could have come from single translations. Therefore candidates for inclusion coming from one, two and three language were manually studied for presence of such vocabulary.

3.3.2 Candidates for deletion

The deletion list contained 644 items. We went through the deletion candidates manually, deleted 137 items and kept 507, guided by the following principles:

We kept candidates for exclusion in the following cases:

- If they belong to some domain of importance to language learners, e.g. "veckodag" (weekday), "väster om" (to the left of)
- if the word is of importance for Swedish traditions/culture, e.g. "midsommar" (midsummer holiday), "fika" (coffee break)
- if the word by its spelling variant/form has not been "recognized" in other lists, e.g. "böra" in the Swedish M2 versus “bör” (non-lemmatized variant) among the inclusion candidates;
- if the word is an item in a deviating presentation form, e.g. in M2 we have “det vill säga (d.v.s.)” which didn’t match “d.v.s.” in translations from other languages
- if the items have to do with the general language, e.g. “kolla” (check), “syfta” (refer)
- if the items have to do with important for Sweden political and social structures, e.g. “konung” (king), “debatartikel” (debate article). Having been involved in the learning/teaching process of Swedish as a Second/Foreign Language we know from experience that newspaper articles are often the core texts in the learning process and therefore words characteristic of them should be kept in the list.
- as far as political terms are concerned, we kept only one item in the “word family”, e.g. of the two exclusion candidates “folkparti”(popular party) and “folkpartist” (adherent of a popular party) only “folkparti” was kept

We deleted the following items:

- words that have functional word classes, e.g. particles, determiners, pronouns
- historical terms, e.g. “stalinistisk”, “bolsjevik”, “marxistisk”, ”koncentrationsläger”
- adverbs if they have been "t"-derived from an adjective present in the M2 list
- some variants of mwe have been removed as a separate entry and added as an example to their headword, e.g. “generellt sett” (in general terms) added as an example to the adverb “generellt”
- political adjectives derived from the names of political parties that are present in the list, e.g. "vänsterpartistisk" (for a leftist party)
- vulgar terms e.g. “jäkel” (bastard, devil); gerundial nouns e.g. “kunnande” (knowing); obvious English loans e.g. “team”, “support”; modern internet-inspired terms “bloggosfär” (blogosphere)
- other words that lacked strong reason for being kept

3.3.3 Candidates for inclusion

Inclusion candidates list comprised 3430 word forms. Of those, 2630 lem-grams have been added.

The 3430 candidates were first checked against a SweWAC lemma list, all possible POS-tags for each item and their WPM frequencies were collected automatically from the same list. As a result a number of items did not match any of the lemmas in the SweWAC and were excluded from the future processing as illegitimate ones. Among the latter ones were non-lemmatized items e.g. dikter (poems, plural), non-existent and misspelled word forms.

We excluded the following items from inclusion candidates:
- items already on the list though in another (lemmatized) form (e.g. "bör": lemma = "böra")
- proper names, e.g. "david", "frankrike"
- participles, e.g. "färgad"
- "t"-adjectives/"t"-adverbs, e.g. “orättvist” if the other form was already on the M2 list
- items not in SAOL (Swedish Academic Word List)
- “-s” verbs that have a corresponding “non-s”-form in M2, e.g. “samlas” vs “samla”
- items ending with "-", e.g. "kärn-"
- items consisting of one letter only: "a", "b", "m"
- items used as translation from only one language, except the items that belong to important for language learners domains, e.g. "aprikos" (apricot), “cykelväg” (bicycle track)
- religious terminology from Arabic
- specific terms from Chinese

The following items were included in the first place:
- Items that have been used in 4-8 languages, if they are among "legitimate" vocabulary, i.e. both in SweWAC lemma-list and in SAOL (Swedish Academic Word List)
- items used in 1 language if they belong to some learner-important domains

- the rest of items (above the threshold of 1) till the total amount of items on the M3-list is 9000 or near it

The resulting candidate list contained a lot of items with multiple POS. Certain POS-tags were erroneously assigned to the items. To minimize the amount of manual work on inclusion candidates, all items that had single POS have been included into M3 without the initial manual control with the prospect of final human check before the delivery of the lists.

Due to the collected SweWAC wpm frequencies, it was possible to place all inclusion candidates relative to the items already on M2-list.

3.3.4 Candidate MWE for inclusion

Out of 530 candidate multiple word expressions, examples were added to 115 headwords, to 44 of those – multiple examples. Altogether 194 mwe were added to the list as examples.

MWE to include:
- to be included into the example column to the headwords (lem-grams) were: phrasal verbs, reflexive verbs, constructions with obligatory prepositions, idiomatic expressions.
- in cases of spelling variants we added variants to the headword entry, e.g. “efter hand” -> “efterhand (el. efter hand)”
- in case the verb is used in combination with a noun e.g. "begå självmord" (commit suicide), we considered adding it to example column for the head noun We avoided inclusion of mwe as new headwords since we did not have the frequency for those and it was against our agreed policy from the start (we didn’t plan having mwe among headwords).

MWE to discard:
- non-idiomatic mwe e.g. “bära in” (bring in), “bära ut” (take out)
- items that are easy to build up of blocks e.g. ”inte kunna” (can’t)
- non-lemmatized forms e.g. “jag kan” (I can);
- items that start with "kvinnlig" (female) e.g. “kvinnlig arbetare” (female employee)
- items otherwise on the inclusion list (“krocka” - word to include; “krocka med” - mwe to include)

3.3.5 Proofreading

Finally, the newly added items (marked as “T2” in the ”Source” column) have been proofread and articles and infinitive markers assigned to nouns and verbs. Prior to proofreading the Swedish M3 list contained 8485 items. As a result of proofreading, the list was reduced to 8425 items.

The division into bands (CEFR levels) was done according to the frequency ranks (ca 1404 words per band):

Band1 (A1) ID 1 - 1404
Band2 (A2) ID 1405 - 2808
Band3 (B1) ID 2809 - 4212
Band4 (B2) ID 4213 - 5616
Band5 (C1) ID 5617 - 7020
Band6 (C2) ID 7021 - 8425

3.4 Universal vs specific vocabulary

3.4.1 Universal vocabulary

Universal vocabulary for all nine partner languages consists of 5 word sets that are symmetrical translations of each other, see table 16.

English	Arabic	Russian	Greek	Norwegian	Swedish	Polish	Italian	Chinese
music	موسيقى	музыка	μουσική	musikk	musik	muzyka	musica	音乐
library	مكتبة	библиотека	βιβλιοθήκη	bibliotek	bibliotek	biblioteka	biblioteca	图书馆
sun	شمس	солнце	ήλιος	sol	sol	słońce	sole	太阳
hospital	مستشفى	больница	νοσοκομείο	sykehus	sjukhus	szpital	ospedale	医院
theory	نظرية	теория	θεωρία	teori	teori	teoria	teoria	理论

Table 16. Symmetric translation sets across all the 9 Kelly languages

A symmetric pair means that the translator of one language, e.g. from English to Swedish has translated let's say "library" as "bibliotek" while the translator from Swedish to English has translated "bibliotek" as "library".

Symmetric set of translations means that (randomly or not) translators between all language pairs have chosen the same variants for the pairs "source word" – "target word" as in the table above.

Examples of non-symmetric translations are the following:

- angå (Swe source) – regard (Eng) versus
- regard (Eng source) – betrakta (Swe)

Some expected words like weekdays, numbers, relative names haven't gained the status of symmetric translation sets. For example the word "bread" is (almost) symmetrically translated, but one of the translators chose to provide an extra variant (synonym) – "corn" and the translator from Norwegian to Arabic provided another variant than the seven other translators to Arabic. The same refers to the word "mother": all translators into Swedish chose the variant "mor" except the Polish 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 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.

The symmetrical sets for 8 languages do not seem to reveal much of a language 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 word. A very interesting example is days of the week that didn't come up among symmetric sets for 9 or 8 languages. One reason for that is Chinese where at least three different names for each weekday are used (depending on the word 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 9 or 8 languages.

Absence of ordinal numerals (one, two, three, etc.) among symmetric sets for 9 and 8 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 hypothesis is that languages from the same language family would share many more symmetric sets, e.g. languages of the Indo-European family (Germanic: English, Norwegian, Swedish; Hellenic: Greek; Romance: Italian; Slavic: Polish, Russian) as opposed to languages coming from other families, e.g. Afro-Asiatic (Semitic: Arabic) and Sino-Tibetan (Sinitic: Chinese).

3.4.2 Common vocabulary for language pairs (Swedish – X language)

The numbers for common vocabulary to the language pairs comprise symmetric pairs for each language combination. Table 17 presents the numbers received for Swedish (that belongs to the Indo-European family, Germanic Subgroup, Northern branch):

Language combination	Number of symmetric pairs for this language combination	Language family/subgroup/branch of the X language
Swedish – Norwegian	3109	Indo-European/Germanic/Northern
Swedish – English	3002	Indo-European/Germanic/Western
Swedish – Italian	2641	Indo-European/Romance
Swedish - Polish	2495	Indo-European/Slavic/Western
Swedish - Russian	2271	Indo-European/Slavic/Eastern
Swedish – Greek	1966	Indo-European/Hellenic
Swedish – Chinese	1123	Sino-Tibetan/Sinitic
Swedish – Arabic	618	Afro-Asiatic/Semitic

Table 17. Shared vocabulary between Swedish and another Kelly language

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 corpus materials the original lists have been derived from as well as approaches to the vocabulary selection.

The highest number of symmetric sets enjoys the pair Swedish-Norwegian: both languages belong to the same family, subgroup and branch (Indo-European/Germanic/Northern). 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 versus 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 differences in the principles of tokenization, lemmatization and vocabulary selection.

3.4.3 Unique vocabulary

Unique vocabulary in this context means that the items listed among the "unique" ones were not used in the translations from other languages to the target language. The lists contain only base forms – no distinction into word classes is kept, which means that certain base forms may be represented more than once in the M3 lists, while gaining only one position in the unique vocabulary list. The following numbers were received for the nine partner languages as presented in table 18:

Language	Number of unique items
Italian	311
English	477
Swedish	501
Greek	590
Norwegian	1119
Polish	1136
Russian	1159
Arabic	2604
Chinese	3051

Table 18. Number of unique items per Kelly language

There are 501 words in the list of unique Swedish words. Of those – surprisingly enough – 15 come from translation "candidate lists for inclusion": four of the words from one language (words belonging to some domain vocabulary), the rest – from two languages. Most probably, other language teams have decided against keeping these words in their final lists while we decided to include them, ironic enough.

The rest of the unique Swedish words represent 118 words marked for domains, while 370 come from the "exclusion list". The latter ones are kept for the reasons described in 3.3.2, among these words are Swedish-specific words like "midsommar", "pingst", "nobelpris", "kvällsmål", "fika" .

4. Statistics and coverage.

4.1 General on vocabulary distribution in the Swedish Kelly-list

The 8425 headwords on the Swedish Kelly-list have been equally assigned to CEFR levels according to their frequency range in the following way:

A1, A2, B1, B2, C1 – 1404 headwords per level

A6 – 1405 headwords

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.
- 2564 headwords come from T2 (translation lists). They constitute 30,4 % of the Kelly-list and cover 1,7% of the SweWAC texts. Approximately 2500 of those items appear in the last two proficiency levels C1 and C2, as shown in table 19.
- 5776 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 1305 and 1377 headwords per level) in the first four CEFR levels, and disappear at all from the last CEFR level C2, see table 19 for all the numbers.

CEFR	Nr of T2 words	SweWAC coverage, %	Nr of SweWAC items	SweWAC coverage, %
1 (A1)	14	0,7	1305	68,9
2 (A2)	27	0,0909	1377	5,3198
3 (B1)	53	0,0882	1351	2,26
4 (B2)	69	0,12	1335	1,16
5 (C1)	996	0,495	408	0,2686
6 (C2)	1405	0,2476	0	0
Total	2564	1,6739	5776	77,98

Table 19. SweWAC coverage by T2 and SweWAC items.

Word class distribution is presented in table 20.

POS	CEFR A1*	CEFR A2*	CEFR B1*	CEFR B2*	CEFR C1*	CEFR C2*	Total (% of Kelly list)*	Coverage, SweWAC
Adjective	170 (2,02 %)	220 (2,61 %)	257 (3,05 %)	240 (2,85 %)	240 (2,85%)	227 (2,69%)	1354 (16,07 %)	6,43%
Adverb	170 (2,02	124 (1,47	116 (1,38	98 (1,16	40 (0,47%	21 (0,25%	569 (6,75%	7,6%

	%)	%)	%)	%)	)	)	)	
Aux.verb	4 (0,05 %)	1 (0,01 %)	-	-	-	-	5 (0,06%)	0,14%
Conjunction	14 (0,17 %)	4 (0,05 %)	1 (0,01 %)	-	-	-	19 (0,23%)	0,41%
Determiner	7 (0,08 %)	1 (0,01 %)	-	2 (0,02 %)	-	-	10 (0,12%)	3,6%
Interjection	5 (0,06 %)	1 (0,01 %)	3 (0,04 %)	5 (0,06 %)	5 (0,06%)	5 (0,06%)	24 (0,28%)	0,1%
Noun	547 (6,49 %)	704 (8,36 %)	762 (9,04 %)	788 (9,35 %)	856 (10,16 %)	950 (11,28 %)	4607 (54,68 %)	14,51%
Numeral	43 (0,51 %)	-	1 (0,01 %)	-	1 (0,01%)	11 (%)	56 (0,66%)	
Participle	-	-	1 (0,01 %)	-	-	-	1 (0,01%)	0,001%
Particle	16 (0,19 %)	4 (0,05 %)	4 (0,05 %)	5 (0,06 %)	-	-	29 (0,34%)	0,45%
Preposition	50 (0,59 %)	23 (0,27 %)	21 (0,25 %)	11 (0,13 %)	2 (0,02%)	1 (0,01%)	108 (1,28%)	11,14%
Pronoun	47 (0,56 %)	5 (0,06 %)	4 (0,05 %)	4 (0,05 %)	1 (0,01%)	-	61 (0,72%)	11,4%
Proper name	11 (0,13 %)	-	1 (0,01 %)	-	1 (0,01%)		13 (0,15%)	
Subjunction	18 (0,21 %)	5 (0,06 %)	3 (0,04 %)	5 (0,06 %)	-	-	31 (0,37%)	1,8%
Verb	302 (3,58 %)	312 (3,70 %)	230 (2,73 %)	246 (2,92 %)	258 (3,06%)	190 (2,26%)	1538 (18,26 %)	16,9%

* the number is given first in absolute count and then in brackets in percent of the total number in the Kelly list

Table 19. Kelly POS distribution in SweWAC

61 pronouns covered 11,4% of SweWAC; 108 prepositions covered 11,14%; whereas 4607 nouns covered only 14,51% compared to 1538 verbs which covered 16,9%. Verbs, pronouns and prepositions therefore appear more “beneficial” to learn than nouns in terms of text coverage, or so it would seem from statistics.

4.2 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 9 genres. SUC has been manually proofread for errors in lemmatization and part-of-speech tagging.

Coverage calculations indicate 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%. 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 in table 21.

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%

Table 21. SweWAC, Parole and SUC coverage in %.

The following coverage numbers we have got per CEFR level, see table 22:

Kelly-words by CEFR	SweWAC coverage, %	Parole coverage, %	SUC coverage, %	Nr of Kelly lemmas
1 (A1)	69,6077	53,35377612	56,9689	1404
2 (A2)	5,41	4,679733157	5,388	1404
3 (B1)	2,34	2,252842572	2,67	1404
4 (B2)	1,28	1,297521605	1,52	1404
5 (C1)	0,7636	0,837440415	1,089	1404
6 (C2)	0,2476	0,328874809	0,537	1405
Inf maker	2,34	1,014368113	1,1	n/a
Proper names	4,9	8,671295408	3,6	n/a
Punctuation	10,7	12,70484044	11,5	n/a
Hapax legomena	0,52	1,185233317	3,19	n/a
Coverage by Kelly-words	79,6489	62,75018868	68,18	8425
Kelly + inf.marker + proper names	96,5	85,14	84,38	n/a

Table 22. SweWAC, Parole and SUC coverage per CEFR level.

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. Reasons might be: (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. 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, 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, 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:

- The zero-matches reflect recent "hot" political events and technological innovations, e.g. *piratparti, svininfluensa, alliansregering, islamist, taliban, reporänta, fildelare, sms* (Eng. *pirate party, swine flu, alliance government, Islamist, Taliban, funding rate, file sharer, sms*);
- 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, webbsida,* (Eng. *to blog, a blogger, blog entry, font size, posting, web browser, website*);
- 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, forums), e.g. *toppen, jävla, tryne* (Eng. *great, damn, snout*);
- 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, brorsdotter* (Eng. *crab, sock, hood, sparrow, sesame, apricot, niece*)
- One more group of zero-matches is constituted by 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 90% of the texts.

5. Lessons learned – summary and conclusions.

5.1 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
5. Validation / post-translation – 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.

5.2 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*.

5.3 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 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’s 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 and some practical information

To summarize, the main characteristics of the Swedish KELLY list are the following:

- it constitutes the most frequent core vocabulary of modern Swedish (as of 2010) derived from a large web-acquired corpus.

- it is based on the objective selection, i.e. human judgment was avoided in favor of objective decisions as much as possible. The word selection has been strictly frequency-based with only 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 primary headword selection has been validated through comparison of basic vocabulary used in eight other partner languages.

- its vocabulary is streamed into CEFR difficulty levels based on frequency and coverage principle.

 The Kelly-list can be downloaded from http://spraakbanken.gu.se/eng/kelly. You are encouraged to make a reference to this report or any other article describing this list (Johansson Kokkinakis & Volodina (2011), Johansson Kokkinakis & Volodina (forthcoming, 2012), Kilgarriff A. et.al. (submitted, 2012), Volodina & Johansson Kokkinakis (accepted, 2012)) if you use the Swedish Kelly-list.

We can conclude by saying that we plan to continue working with the Swedish KELLY list in the future. The way it has been compiled, it addresses a number of target user groups, including language teachers, test producers, lexicographers, comparative linguists, computational linguists, etc. 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

Allén, Sture (2006). *Libers lilla ordlista*. Liber, Sweden.

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

Allén, Sture (2002). *Våra viktiga ord*. Liber, Sweden.

Andréasson Maia, Borin Lars, Merkel Magnus (2008). *Habeas Corpus: A survey for SNK - a Swedish national corpus.* University of Gothenburg, Sweden.

Borin L., M. Forsberg and L. Lönngren (2008). The hunting of the BLARK - SALDO, a freely available lexical database for Swedish language technology. Resourceful language technology. *Festschrift in honor of Anna Sågvall Hein*, ed. by Joakim Nivre, Mats Dahllöf and Beata Megyesi. Acta Universitatis Upsaliensis: Studia Linguistica Upsaliensia 7. 21-32.

Council of Europe. (2001). *The Common European Framework of Reference for Languages*. Retrieved from http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf Access date: March 2nd 2010

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*. <http://stp.lingfil.uu.se/~evafo/resources/basevocpool/>

Gardner, D. (2007). Validating the Construct of Word in Applied Corpus-based Vocabulary Research: A Critical Survey. *Applied Linguistics* 28/2, p.241-265.

Gellerstam, Martin (1978). *Välja sina ord*. Rapporter from Språkdata, Göteborgs Universitet, Sweden.

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.

Johansson Kokkinakis, S. and Volodina, E. (forthcoming 2012). En svensk ordlista för språkinlärning - ett korpusbaserat angreppssätt i EU-projektet Kelly. *NFL 2011*, Lund, Sweden.

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

Kilgariff Adam, Rychly Pavel, Smrz Pavel, Tugwell David (2004). The Sketch Engine. *Proc EURALEX 2004*, Lorient, France.

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

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, Sweden.

Köhler, Per Olof & Messelius, Ulla (2001). *Natur och Kulturs Svenska Ordbok*. Natur och Kultur Läromedel, Sweden.

Lexin (2006). *Svenska Ord: med uttal och förklaringar*. Språkrådet, Sweden.

Ma, Q. and Kelly, P. (2006). Computer-Assisted Vocabulary Learning: Design and Evaluation. *Computer-Assisted Language Learning* 19, 15-45.

Nation, P. and Waring, R. (1997). Vocabulary Size, Text Coverage and Word Lists. *Vocabulary: Description, Acquisition and Pedagogy*, 6-19.

Nygren, Håkan (2010). *Svensk skolordlista*. Norstedt, Sweden.

Praktisk Svensk Ordlista (1993). Svenska språknämnden och Svenska akademien.

Read, J. (2000). *Assessing Vocabulary*. Cambridge University Press.

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

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

GU-ISS, Forskningsrapporter från Institutionen för svenska språket, är en oregelbundet utkommande serie, som i enkel form möjliggör spridning av institutionens skriftliga produktion. Det främsta syftet med serien är att fungera som en kanal för preliminära texter som kan bearbetas vidare för en slutgiltig publicering. Varje enskild författare ansvarar för sitt bidrag.

GU-ISS, Research reports from the Department of Swedish, is an irregular report series intended as a rapid preliminary publication forum for research results which may later be published in fuller form elsewhere. The sole responsibility for the content and form of each text rests with its author.

Forskningsrapporter från institutionen för svenska språket, Göteborgs universitet
Research Reports from the Department of Swedish

ISSN 1401-5919

www.svenska.gu.se/publikationer/GU-ISS