A Swedish Core Vocabulary for Machine Translation

Annette Östling
Uppsala University

Abstract

The reasons for establishing a new Swedish core vocabulary are presented, and the steps taken in its establishment are described. Some conclusions are drawn as for the usefulness of frequency lists in this respect. The impact on the frequency lists of the nature of the corpus is illustrated. The necessity of introducing phrases in the core vocabulary is pointed out. The information to be associated with the entries to meet the requirements of the translation process is looked upon in the light of the definitions in a monolingual dictionary.

Introduction

The project *Multilingual Support for Translation and Writing* carried out at the Department of Linguistics/Computational Linguistics at Uppsala University aims at a computer tool for the processes of translation and writing, with professional translators as the main user group in view. The translation process is to be monodirectional with Swedish as the source language and English, German and French as the targets. Apart from providing possibilities for mechanical translation of parts of a document specified by the user, the purpose of the support is also to provide functions for selective dictionary look-up, allowing the user to specify the kind of information wanted about a requested word or phrase. The dictionary is to be organized as a lexical database, made up of monolingual databases for Swedish, English, German and French. The links between the Swedish database and the target language ones will be the equivalence relations between the translation units in a language pair. A translation unit is defined as the smallest possible unit in the source language that can be substituted by an equivalent in the target language, on the semantic level as well as on the stylistic level. It can be a word, a phrase or a larger segment of the text. It is a well-known fact that one-to-one correspondences between the lexical units, words or phrases of two languages are rare. Contrastive lexical studies, investigating the equivalents and their relations are central in translation theory, and hence are important also for machine translation. In Wikholm (1991) an account is given of the types of lexical differences that exist between languages and different types of equivalence relations are discussed, along with the implications for machine translation.

Within the monolingual databases of the LDB, there will be a distinction between a general, permanently stored vocabulary and domain-specific parts. The general part, the core dictionary, will cover high-frequency, common language words and phrases, whereas the users will incorporate the domain- or text-specific vocabularies in the LDB in accordance with their needs.

It is important to have in mind that the LDB is to be consulted by a human translator/writer as well as by the machine translation component of the support and must fulfill the demands of both kinds of users.

Since the primary units of the database are the translation units, the Swedish core dictionary must be structured in such a way that the establishment of translation relations to the target languages is possible. High-frequency words are often manyfold ambiguous, morphologically and semantically. For example, the English word 'in' can be either a preposition, an adverb or a noun, and there are obviously no one-to-one correspondences between 'in', 'do' and 'make' and their Swedish equivalents. In phrases the ambiguity can be resolved: 'in' is disambiguated as a preposition when used in the phrase 'in accordance with'. This fact will be taken advantage of by introducing phrases as an important part of the core vocabulary, and is one step in the development of a core vocabulary of translation units.

In this project, Swedish is the source language. Therefore, in a first phase, the Swedish core vocabulary will be established. Then translation links to the equivalents in English, German and French will be determined.

Is there a need for a new Swedish core vocabulary?

A first Swedish basic vocabulary was identified within the project *Nusvenskfrekvensordbok (NFO), 'Frequency Dictionary of Present-Day Swedish'* (Allén et al. 1970). The NFO basic vocabulary was derived from a corpus of Swedish newspaper text from 1965, consisting of one million running words. The basic vocabulary comprises 10,000 lemmatized graphic words[1]. Thus no compound expressions such as phrases or phrasal verbs are recognized as units in the NFO basic vocabulary, the basic reason for the decision to determine a new core vocabulary within this project. The size is another important issue: is 10,000 graphic words a reasonable size for a core vocabulary of the kind needed for a tool for translation and writing? Since the basic vocabulary of NFO is quite big, it comprises a high proportion of content words. It is in the nature of content words to reflect the topics treated in a text, and some of the current issues of 1965 newspaper articles are no longer of the same importance (the Vietnam war, for instance), and many new topics and phenomena have arisen since then, such as the concern for the environment, the media explosion with video and satellite television, etc. As shown below, even words in the frequency top can be quite domain-dependent. Thus part of the NFO basic vocabulary may be, in this respect, somewhat outdated.

The NFO basic vocabulary was determined as a general frequency top list, without any special use in mind. The core vocabulary aimed at in this project is to be a tool for all the users in view, regardless of their various fields of application. Hence one of the main criteria for the choice of which content words to be incorporated must be their neutrality with respect to domain, or, put in another way, the likelihood for them being used in different domains. It follows that it is natural to look not only at frequency lists of newspaper vocabulary, but also to compare with frequency lists of other types of texts. Lehmann (1991) describes the comparison of five word lists from very different domains, and shows that the common vocabulary is very small indeed. He also points out that the definition of a representative corpus, and hence a representative vocabulary, must be made with the research goals pursued in mind. It is quite clear that a core vocabulary for the purpose of being used as a translation tool as the one described here is not the same thing as a basic vocbulary for language learning, for instance. A dictionary for this latter purpose must cover the vocabulary and language functions for every-day life, since it aims at guiding the language student to what is the "threshold level" of another language, something which is not the same thing as domain-neutrality.

How can a core vocabulary be determined?

The need for core vocabularies for the purpose of natural language processing is pointed out in Lehmann (1991), who also describes the establishment of a German core dictionary.

The new Swedish core dictionary for the translation and writing tool being worked out within this project is based on the total of the newspaper material available via Språkbanken ('The Language Bank') at Språkdata, University of Gothenburg (Gellerstam 1989). In all, this material comprises approximately seven million current words from 1965, 1976 and 1987, the words from 1987 constituting the largest part: five million words. Comparisons have been made with the novel corpus from Språkbanken. This corpus consists of two parts, totalling 9.6 million words^2: novels published in 1976 and 1977, comprising 5.6 million running words, and novels published in 1980 and 1981, four million running words.

The first step in establishing the core vocabulary was to make a morphological analysis of all the word forms from the Språkbanken newspaper corpus, henceforth NWP, with a frequency of 160 or more (the 2,926 most common word forms). The morphological analysis was performed with the help of Uppsala Chart Processor (UCP) (Sågvall Hein 1987), an inflectional grammar of Swedish in the UCP formalism (Sågvall Hein, forthcoming) and a stem dictionary of about 60,000 entries (Sågvall Hein & Sjögreen 1991). The stem dictionary is generated from *Svensk Ordbok, 'A Dictionary of Swedish'* (1986) and thus covers the same vocabulary. It follows that the morphological analysis results in the same lemma distinctions as the ones made in *Svensk Ordbok.*

Homography is a very common phenomenon in Swedish, not only among high-frequency function words, and therefore the number of possible lemmas assigned to a word form by the morphological analyzer is often > 1. The frequency 160 was chosen ad hoc as a starting point, but proved to be a good guess (see below). The result of this parsing was a list of all the possible lemmas for each word form according to *Svensk Ordbok,* as shown by the following example. Here, the graphic word *perfekt* is given one noun analysis (the grammatical term 'perfect') and one adjective analysis ('perfect'):

```
perfekt: (freq: 311)
2 parses, 9 vertices.
        PERFEKT :
        (* =    (LEM = PERFEKT2.NN          (* =    (LEM = PERFEKT1.AV
                INFL = PATTERN.GRYN                 INFL = PATTERN.BLEK
                DIC.STEM = PERFEKT                  DIC.STEM = PERFEKT
                WORD.CAT = NOUN                     WORD.CAT = ADJ
                GENDER = NEUTR                      COMP = POS
                FORM = INDEF                        NUMB = SING
                CASE = BASIC))                      FORM = INDEF))
```

Since this is a pursuit of a *core* vocabulary, all lemmas marked for a certain level of style or a specific domain were excluded from the list. Furthermore, the lemmas that only occur in compounds were given a special code.

The next step was to compare this list of graphic words tagged with their possible lemmas with a lemmatized frequency list in order to exclude unlikely lemma attributions. Thus a comparison, much like a manual probabilistic tagging, was made with *Nusvenskfrekvensordbok,* and the lemmas having a NFO frequency of 10 or less (i.e. the lemmas in question are not among the 7,150 most common ones) were excluded from the list. For instance, the noun analysis of *perfekt* was among the ones sorted out. Setting the limit as low as f ≤ 10 may seem an unnecessary precaution, but in this early sorting out precaution is necessary in order not to rule out lemmas that are rare only in

NFO. Of course lemmas that should not have been excluded may still have been sorted out by mistake. Testing against the texts of the potential users will be necessary in order to find inconsistencies and to refine the vocabulary so far defined.

In order to study the frequency of the different parts of speech more in detail, the list of graphic words tagged with their possible lemmas was divided into 14 segments of 200 word forms each. Segment 1 consists of the 200 most common word forms, segment 2 comprises the next 200 and so forth. The last 126 graphic words were thereby left out, something which does not seem to be of much importance (see below). The following table shows the frequency span for the graphic words included in each segment (the corpus size is seven million current words):

Graphic word frequency

Segment 1	201670-2510	Segment 8	352-306
Segment 2	2497-1194	Segment 9	306-269
Segment 3	1192-820	Segment 10	269-243
Segment 4	818-617	Segment 11	243-215
Segment 5	616-497	Segment 12	215-196
Segment 6	497-412	Segment 13	196-182
Segment 7	412-352	Segment 14	182-167

The table below gives an account of the number of possible lemmas of every word class found in each segment. The lemmas with a NFO frequency of 10 or less have been excluded from the table. S1 etc. refers to segment number[3].

	Noun	Verb	Adj	Adv	Pron	Prep	Con	Subj
S1	26	55	30	67	39	29	19	1
S2	61	48	48	51	15	9	3	1
S3	90	50	35	32	9	3	1	1
S4	106	62	41	11	2	4	1	3
S5	89	60	54	20	1	2	0	1
S6	100	58	46	11	2	3	3	0
S7	99	52	45	10	1	0	0	1
S8	109	66	33	12	2	2	0	0
S9	108	66	39	11	1	2	0	0
S10	112	54	43	10	0	1	1	2
S11	91	76	38	6	1	2	0	0
S12	117	64	35	4	1	1	0	0
S13	108	55	36	10	1	1	0	0
S14	103	71	32	8	3	1	0	0
Σ	1406	837	555	263	78	60	28	10

NB that the number of possible lemmas in each segment is over 200. This is due to the fact that many graphic words have more than one lemma attribution. For instance, many graphic words in segment 1 have been analyzed as being both possible prepositions and adverbs (i.e. in this case verb particles). *får*, on the other hand, an extremely common verb form meaning, among other things, 'may', 'should' or 'has', is homograph with the noun meaning 'sheep'. Since the frequency of the noun lemma *får* in the NFO frequency list is less than 10, this lemma attribution has been discarded.

The segmentation procedure proved fruitful for many decisions, especially concerning the function words, since it made it possible to see in which segments they start disappearing. It also permitted a close study of where on the frequency scale the domain-specific content words begin outnumbering the more general ones, although this study still could not be decisive for which content words to incorporate in the core vocabulary. The segmentation made it evident, though, that the distribution of the content word lemmas is fairly even throughout the segments (apart from the first two for the nouns). It is thereby clear that the content words give their color even to the frequency top, something to have in mind when discussing the representativity of a corpus for the goals in view. Below follows a brief discussion of each word class. For a list of all the entries proposed for the core vocabulary, see Östling (forthcoming).

Function words

Prepositions

All prepositions found in this 2,800 word form top list seem to qualify for inclusion in the core vocabulary. It is evident that after f 2,000, i.e. segment 10, the number of "new" prepositions is very low, something which permits the preliminary conclusion that the prepositions found here really are the core ones in Swedish – it seems unlikely that some important prepositions appear below this frequency. Testing against user texts and other corpora would of course corroborate or falsify this hypothesis.

Conjunctions

It is interesting to see that the pattern for the prepositions is repeated, and is even clearer, for the conjunctions. Below f 2,000 no conjunctions appear, which supports the hypothesis that neither for this part of speech will it be necessary to go further down the frequency list. The conjunctions among the top 2,000 can thus tentatively be claimed to be the conjunctions of the core vocabulary.

Subjunctions

The subjunctions are very few, but the pattern remains the same as for the two parts of speech above: there are no subjunctions below f 2,000, and the list of subjunctions established from the frequency top here is preliminarily established as the list of the core vocabulary.

Pronouns

Below segment 3, i.e. below f 600, only a handful of pronouns are to be found. A closer study, however, shows that some personal and possessive pronouns are missing. The use of these pronouns is highly dependent on the communicative function of the text (Lehmann 1991), and has nothing to do with the topics treated. It is evident that in a core vocabulary of the kind needed here the pronouns of all six grammatical persons should be included. In order to find them all in this frequency list, it would be necessary to go beyond the top 2,800, but this would be a superfluous task, since it is easy to check this consistency manually in the existing list of pronouns. (Of course it is not uninteresting to see where on the frequency list the missing ones appear – this could be of importance for core vocabularies proposed for other types of use, as for instance language learning).

Adverbs

Many Swedish adverbs are formed on an adjective stem, the ending -t signalling the adverbial function. This ending, however, is also the neutral gender ending of Swedish adjectives, and this latter use is the only one recognized by the UCP parser. Thereby no adverbs of this open-class kind are included among the core vocabulary adverbs so far, something which of course has to be remedied in due course of time.

The majority of the adverbs among the top 200 are commonly used as verb particles. This is true, for instance, for *in* (*komma in* – 'come in') and *över* (*ta över* – 'take over'). The number of "new" adverbs is diminishing throughout the segments, but the segmentation gives no evidence for the conclusion that basic adverbs will not be found below f 2,800. Thus the frequency list has to be explored beyond this limit, and the procedure described for the content words (see below) will have to be used.

Content words

A quick glance at the frequency list reveals the fact that surprisingly many content words with a high frequency are somehow specialized and domain-specific. Some noun examples:

	Freq	Segment
regeringen ('the government')	2,813	1
matchen ('the match')	1,371	2
kommunen ('the local authorities')	1,284	2
VM ('world championship')	930	3

A newspaper corpus consists of articles of varied topics, but it is clear that some topics are more frequent than others, that the current topics vary with time, and that some topics hardly ever are treated in newspapers. In spite of this, newspaper corpora are often judged to give a reliable cross-section of contemporary language. As a somewhat provocative contribution to this debate, the nouns among the top 2,800 in the NWP that are not to be found in the same segment of the novel corpus are most illustrative. Here follows a list of some of them, randomly chosen:

aids aids	*bolag* company	*final* final	*motståndare* adversary
aktie share	*budget* budget	*försvar* defence	*personal* staff
argument argument	*daghem* day-nursery	*kongress* congress	*resurs* resource
befolkning population	*debatt* debate	*kostnad* expense	*satsning* venture
bidrag allowance	*expert* expert	*lăsare* reader	*syfte* purpose

It is, in our opinion, quite clear that these words are typical of newspaper texts. The presence of *aids* in the list also shows that this corpus must be a fairly recent one. Intuitively it is easy to conclude that the words are not likely to be used to the same extent in fiction, and cannot be labelled domain-neutral. If, on the other hand, we take a closer look at some of the novel nouns that are not to be found among the top 2,800 in the NWP, we find another type of nouns, intuitively felt to belong to the world of fiction and almost forming a poem just by themselves:

blick look	*frukost* breakfasti	*idiot* diot	*mŏrker* darkness
disk dishes	*gryning* dawn	*klănning* dress	*skratt* laughter
drŏm dream	*gråt* tears	*kŏksbord* kitchen table	*smărta* pain
famn arms	*hemlighet* secret	*lăngtan* longing	*suck* sigh
fest party	*hud* skin	*måne* moon	*săng* bed

This latter list could have been made even more "fiction biased", comprising only parts of the body and emotions, for instance.

This clearly shows the importance of studying corpora of different nature, if some degree of domain-neutrality is to be achieved. "Domain-neutrality" may be a notion which does not exist in reality – no word is neutral when it is used, but on the contrary gets some of its significance from the context. Domain-neutrality is thus something which can only be aimed at and probably never achieved. In order to approach this goal, the common procedure for the nouns, adjectives and verbs was to determine the intersection set between the two corpora of the top 2,800 of the parts of speech of the content words. Thereby the words chosen for the core vocabulary are used frequently in newspaper texts as well as in fiction, a clear indication that they are commonly used in domains of different types.

Nouns

It resulted from the study of the NWP that the raw frequency figures were of no great help when it came to deciding where to draw the limit for the nouns to be included in the core vocabulary: domain-specific nouns as well as more general ones occur in each segment. One important conclusion from this study was that it is not possible just to draw a limit in a frequency list, but that frequency studies can merely be one part of the material needed for conclusions concerning the content words. Semantic classification and "common sense" also have to play important roles.

The intersection with the novel corpus resulted in a list with 472 noun lemmas (many lemmas were represented by more than one graphic word). This list was then checked against the lemmatized NFO list, and some unlikely lemma attributions could be deleted. The nouns were then grouped semantically: family, professions, time & season, nature & weather, communications, body, institutions & community, food & drink, house & home, directions, daily things, attitudes, materials, measurements, arts. These groups captured a good deal of the nouns, and permitted the spotting of inconsistencies in the frequency lists. It was found, for instance, that all the weekdays except *lördag* and *söndag* ('Saturday' and 'Sunday') were missing. This grouping was most useful in order to find missing hyperonyms and missing and superfluous hyponyms, and was thereby of great help in the creation of more homogeneous groups. Quite a few hyperonyms were among the nouns added to the list on its way towards a higher degree of generality. In spite of the usefulness of the intersection with the novel corpus and the semantic grouping, it has to be pointed out that common sense still must play an important role for many decisions.

The "final" list consists of 400 noun lemmas, but it is very likely that it has to be further restrained when tested against user texts and other corpora.

Adjectives

The same procedure as for the nouns was carried out for the adjectives. The intersection with the possible adjective lemmas in the novel corpus was established, giving a list of 243 adjective lemmas. The list was checked against the lemmatized NFO material to further exclude rare and unlikely adjective lemmas. A rough semantic grouping was then made to check that no elementary adjectives or basic colors were missing. In this way some inconsistencies were found, and after these procedures, plus the important use of common sense, the ameliorated list today contains 172 lemmas.

Verbs

The verbs have been less thoroughly studied than the other parts of speech. This is due to the fact that no study yet has been made taking into account the very common phrasal verbs (*komma in, ta över*). The procedure followed so far is equivalent with the one used for the adjectives and the nouns. The intersection group with the novel corpus comprises 271 verb lemmas, but these have not yet been checked against the NFO material, neither has a semantic classification been performed, and the list is also in these respects less worked out compared to the others. The very preliminary verb list established so far comprises 265 simple verb lemmas.

Phrases in the core vocabulary

As already pointed out, ambiguity is a very common phenomenon among high-frequency function words. A step towards the establishment of translation units between Swedish and the target languages is the introduction of phrases in the core vocabulary. There are lexicalized phrases of many types: invariable, continuous phrases, an example of which is *i enlighet med*, 'in accordance with', inflectible phrases including phrasal verbs such as *stänga av*, 'turn off' and discontinuous ones, as for example *antingen ... eller*, 'either ... or'. The phrases focussed on so far in the project are the invariable continuous ones, the *functional core phrases*. Criteria for the determination of this type of phrases have been worked out (Sågvall Hein, Östling & Wikholm 1990) and include the following points:

- a functional core phrase should form a syntactically motivated unit. A phrase can not include an element which is part of a previous or following constituent. It follows that the phrase must be continuous.
- it must have a specific grammatical function: prepositional, adverbial, adnominal or conjunctional.
- it should be semantically neutral. If the phrase contains a nominal element, this has to be non-referring.
- it should disambiguate one or more elements of the phrase with regard to meaning or part of speech (see the example mentioned above concerning 'in').

A pilot study showed that the structures which are potential functional core phrase candidates are limited in number, and for Swedish the following four structures are the only possible ones:

preposition + noun + preposition
i enlighet med, 'in accordance with'
inom ramen för, 'within the framework of'

preposition + adjective/pronoun + noun
på goda grunder, 'for excellent reasons'
i första hand, 'in the first place'

preposition + noun
med flit, 'on purpose'
i år, 'this year'

adverb + adverb
snarast möjligt, 'as soon as possible'
inte minst, 'not least'

An important source for the establishment of the Swedish functional core phrases is the *Frequency Dictionary of Present-Day Swedish, Volume 3, Collocations* (Allén et al. 1975). It proved that less than half the number of the collocations of the structures above could be included in the core vocabulary. The non-fulfillment of the criteria was often due either to specific reference of the noun or the adjective, or to the preposition being dependent on a preceding verb, thus invalidating the criterion of a syntactically motivated unit ('learn from | this development', for example, in which the preposition is dependent on 'learn'. In this particular example, the noun also has specific reference). The list of functional core phrases is being worked out, and will include some 1,000 entries (Östling & Wikholm, forthcoming).

In a later phase of the project, inflecting phrases, such as phrasal verbs and temporal expressions will be included in the core vocabulary, along with some common abbreviations.

What information is to be associated with the entries in the Swedish database?

The information associated with each entry is of crucial importance for the functionality of the database, for the human user as well as for the machine translation component. The ideas presented below are part of our working concept, and have not yet been implemented.

In general terms, it is a quite simple task to determine which function words that are to be included in the core vocabulary, but it is most difficult to determine what information is to be associated with them for a nicely functioning translation and writing support. For the content words, the opposite is true: it is most difficult to decide on which ones fulfill the criterion of being sufficiently domain-neutral to be included in the core vocabulary, but it is an easier task, although far from trivial, to decide on what information is to be associated with them.

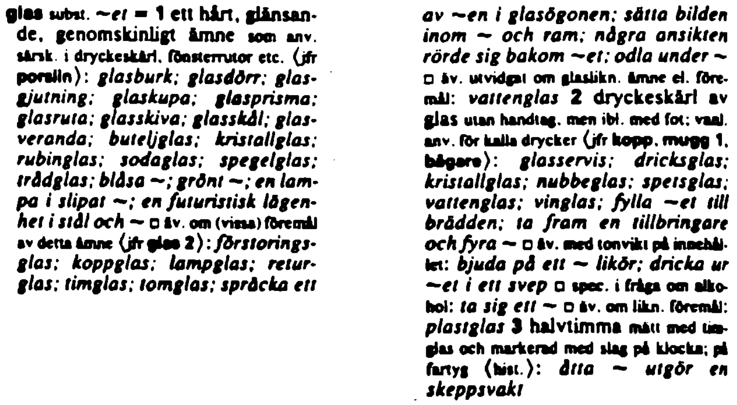
Morphological information is to be an essential part of each entry, and should be consistent with the output of the UCP parser (Sågvall Hein 1987), as exemplified above for *perfekt*. The morphological analysis provides the lemma, LEM, in accordance with the classification in *Svensk Ordbok*. The part of speech of the lemma is given after the punctuation mark, in the example nn, 'noun', and av, 'adjective'. INFL gives information about which inflectional pattern the lemma belongs to. DIC.STEM specifies the stem, since there is sometimes a difference in Swedish between the lemma and its stem. The analysis also provides morphological information specific to each part of speech, in the case of nouns gender, number, form and case.

Each lemma is associated with a list of lexemes, classified according to *Svensk Ordbok* (Sågvall Hein 1988). Thus the lemma glas.nn ('glass') is linked to its lexemes in the following way:

```
(GLAS.NN        (LEX GLAS.NN.1)
                (LEX2 GLAS.NN.2)
                (LEX3 GLAS.NN.3))
```

One type of semantic information in the database is thereby the definitions of the lexemes.

The entry for glas is a good starting point for a discussion of the information to be associated with the noun entries:



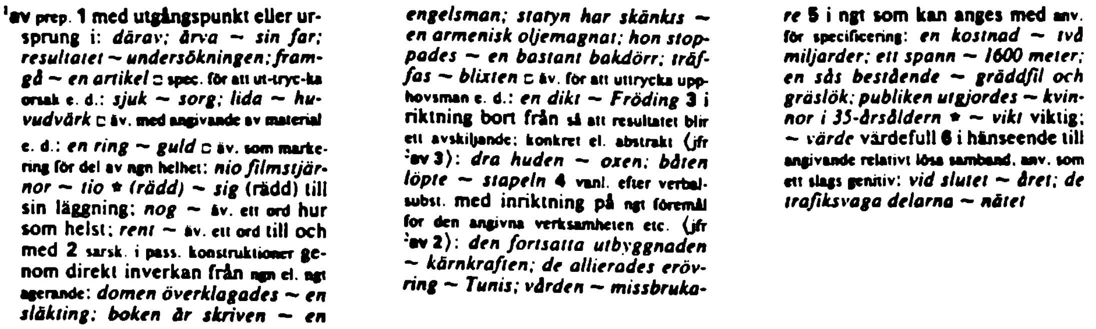
glas subst. *~et* = **1** ett hårt, glänsande, genomskinligt ämne som anv. särsk. i dryckeskärl, fönsterrutor etc. (jfr **porslin**): *glasburk; glasdörr; glasgjutning; glaskupa; glasprisma; glasruta; glasskiva; glasskål; glasveranda; buteljglas; kristallglas; rubinglas; sodaglas; spegelglas; trådglas; blåsa ~; grönt ~; en lampa i slipat ~; en futuristisk lägenhet i stål och ~* □ äv. om (vissa) föremål av detta ämne (jfr **glas 2**): *förstoringsglas; koppglas; lampglas; returglas; timglas; tomglas; spräcka ett av ~en i glasögonen; sätta bilden inom ~ och ram; några ansikten rörde sig bakom ~et; odla under ~* □ äv. utvidgat om glaslikn. ämne el. föremål: *vattenglas* **2** dryckeskärl av glas utan handtag, men ibl. med fot; vanl. anv. för kalla drycker (jfr **kopp, mugg 1, bägare**): *glasservis; dricksglas; kristallglas; nubbeglas; spetsglas; vattenglas; vinglas; fylla ~et till brädden; ta fram en tillbringare och fyra ~* □ äv. med tonvikt på innehållet: *bjuda på ett ~ likör; dricka ur ~et i ett svep* □ spec. i fråga om alkohol: *ta sig ett ~* □ äv. om likn. föremål: *plastglas* **3** halvtimma mätt med timglas och markerad med slag på klocka; på fartyg (hist.): *åtta ~ utgör en skeppsvakt*

The first two lexemes of *glas* are to be included in the core dictionary, whereas the third one has the comment <hist.>, 'historical', and hence must be discarded, since its usage is limited to a certain domain. The distinction between the two first lexemes is crucial from a translational point of view: glas.nn1 is the substance, whereas glas.nn2 is the vessel used for beverages. This information can be extracted from the definitions in the following way: glas.nn1 is defined as an *ämne*, 'substance', and with a semantic parsing, an ISA-link can be established (Vossen et al. 1989, Hagman 1991) to the hyperonym, which is thereby made evident. For glas.nn2 the ISA-link is established to *dryckeskärl*, 'vessel used for beverages', the hyperonym in question. Especially useful for the writing-process is information concerning hyponyms. These are enumerated after the meaning descriptions in the definitions in *Svensk Ordbok*: for glas.nn2 "*dricksglas; kristallglas;*" ('drinking glass', 'crystal glass') etc. Synonyms can also be of use, for the translation as well as the writing process, and this relation can be extracted from the meaning description in the following way: if the ISA-attribute has a numerus value matching the lemma and has no modifiers, it is considered a synonym of the lemma (Hagman 1991). Sometimes synonyms are also explicitely listed in *Svensk Ordbok* after the meaning description, and are thus easily retrieved.

Syntagmatic information, such as the use of the lexeme in expressions, idioms and phrases, is most important in order to achieve idiomatic translations. *Svensk Ordbok* gives quite a few examples of the lexemes in use, and some information of this kind can thereby be accessed from the definitions. It is crucial for the functioning of the LDB that the syntagmatic information be looked upon and determined in the light of what the translation units between the languages are. Since the translation units are not always the same between the language pairs, the syntagmatic information has to be specified with regard to the different target languages.

As for the contrastive information, the reference from the Swedish lexeme to its equivalent(s) in the target languages is the only link between them. glas.nn1 has one equivalent in English, 'glass (mass noun)', and the equivalent of glas.nn2 is 'glass (countable)'. The lexeme distinction reflects the different translation equivalents in this case. Often the use of a word in a phrase affects the translation, and equivalents have to established between these larger unities in the definitions, thereby overriding the translation equivalent of the simple lexeme.

The entry for the preposition *av* illustrates the situation for the function words:



¹av prep. 1 med utgångspunkt eller ursprung i: *därav; ärva ~ sin far; resultatet ~ undersökningen; framgå ~ en artikel* ◻ spec. för att ut-tryc-ka orsak e. d.: *sjuk ~ sorg; lida ~ huvudvärk* ◻ äv. med angivande av material e. d.: *en ring ~ guld* ◻ äv. som markering för del av ngn helhet: *nio filmstjärnor ~ tio* ✱ *(rädd) ~ sig* (rädd) till sin läggning: *nog ~* äv. ett ord hur som helst: *rent ~* äv. ett ord till och med 2 särsk. i pass. konstruktioner genom direkt inverkan från ngn el. ngt agerande: *domen överklagades ~ en släkting; boken är skriven ~ en engelsman; statyn har skänkts ~ en armenisk oljemagnat; hon stoppades ~ en bastant bakdörr; träffas ~ blixten* ◻ äv. för att uttrycka upphovsman e. d.: *en dikt ~ Fröding* 3 i riktning bort från så att resultatet blir ett avskiljande: konkret el. abstrakt (jfr ²av 3): *dra huden ~ oxen; båten löpte ~ stapeln* 4 vanl. efter verbalsubst. med inriktning på ngt föremål för den angivna verksamheten etc. (jfr ²av 2): *den fortsatta utbyggnaden ~ kärnkraften; de allierades erövring ~ Tunis; vården ~ missbrukare* 5 i ngt som kan anges med anv. för specificering: *en kostnad ~ två miljarder; ett spann ~ 1600 meter; en sås bestående ~ gräddfil och gräslök; publiken utgjordes ~ kvinnor i 35-årsåldern* ✱ *~ vikt* viktig; *~ värde* värdefull 6 i hänseende till angivande relativt lösa samband, anv. som ett slags genitiv: *vid slutet ~ året; de trafiksvaga delarna ~ nätet*

The lemma av1.pp has six lexemes. It is not always evident that the division in lexemes coincides with the different translation equivalents. The lexeme av1.pp1 has at least two equivalents in English, often dependent on the preceding noun or verb: *ärva av* – 'inherit from', *resultatet av* – 'the result of', *rent av* – 'even'. These few examples clearly show the need for the establishment of translation equivalents between phrases. av1.pp2 refers to a grammatical construction, the passive, and hence this lexeme should be more uniform in its behaviour in the target languages. av1.pp5 exemplifies the difficulties in establishing the translation equivalents with a Romance language like French, whose constructions are often quite different from those of the two Germanic target languages: *ett spann av 1600 meter* ('a span of 1,600 metres') – 'une arche longue de 1600 mètres'. In cases like this, the translation unit would be the whole phrase, impossible to translate without knowledge of the nature of the head noun – is the proper adjective high, tall, long? To conclude, the lexeme distinction of the function words is not always of help when it comes to establishing translation equivalents. The syntagmatic information is often decisive for the establishment of the translation equivalents to the target languages, and has to be determined contrastively. The introduction of phrases as translation equivalents is necessary for the database to be of help for the human translator and for the functioning of the machine translation component.

Conclusion

It proved that the mere use of a frequency list based on newspaper material was not sufficient for deciding which content words to include in the Swedish core vocabulary. Since a more domain-neutral set of content words was desirable for the purpose in view, a comparison with novel corpus frequency lists was undertaken, and semantic classifications also proved necessary. For the function words, on the other hand, the segmentation of the frequency top of the newspaper material seems to be a sufficient procedure to establish which ones to be incorporated in the core vocabulary. The ambiguity, especially prevalent among the high-frequency function words, calls for the introduction of phrases into the core vocabulary, thus making possible the establishment of equivalence links between the translation units in Swedish and the target languages. As for the information to be associated with the entries in the database, much can be extracted from the definitions in *Svensk Ordbok* through a semantic parsing, especially as regards the content words.

Notes

1. A graphic word is defined as a segment delimited by blanks, punctuation marks or a line feed.
2. The novel is thus bigger than the newspaper corpus, something which makes a direct comparison between the figures impossible. Since the novel corpus is used only to balance, or neutralize, the material, mathematical accuracy is not of primary importance, and no exact calculations have been made.
3. Numerals, interjections, proper nouns and the infinitive marker have been excluded from the frequency study of the following reasons: there is only one infinitive marker, and it is obvious that is should be included in the core vocabulary. Proper nouns are domain-specific, and are thus generally disqualified for inclusion, possibly with the exception of the country names and nationality adjectives referring to the countries in which the database languages are the mother tongues. As for the interjections, *ja, nej* and *tack* ('yes', 'no' and 'thank you') are evident candidates for incorporations in the core vocabulary. The basic numerals must of course also be included.

References

Allén, S. et al. 1970. *Nusvensk frekvensordbok.* 1. Graford. Homografkomponenter. [Frequency Dictionary of Present-Day Swedish. 1. Graphic Words. Homograph Components]. Stockholm.

Allén, S. et al. 1975. *Nusvensk frekvensordbok.* 3. Ordförbindelser. [Frequency Dictionary of Present-Day Swedish. 3. Collocations]. Stockholm.

Gellerstam, M. 1989. The Language Bank. Department of Computational Linguistics. University of Gothenburg.

Hagman, J. 1991. Common and Odd Relations in Italian Dictionaries and Their Treatment in Taxonomy Building. ACQUILEX Working Paper, Istituto di Linguistica Computazionale del CNR, Università di Pisa.

Lehmann, H. 1991. Towards a Core Vocabulary for a Natural Language System. Proceedings of the 5th ACL Conference. Berlin.

Östling, A. A Proposal for a Swedish Core Vocabulary. Simple Units. Department of Linguistics, Uppsala University. Forthcoming.

Östling, A. & Wikholm, E. A Dictionary of Functional Core Phrases. Swedish, English, German and French. Department of Linguistics, Uppsala University. Forthcoming.

Sågvall Hein, A. 1987. Parsing by Means of Uppsala Chart Processor. (UCP). In: Bolc, L. (ed). *Natural Language Parsing Systems*. Springer Verlag.

Sågvall Hein, A. 1988. Towards a Comprehensive Swedish Parsing Dictionary. In: *Studies in Computer-Aided Lexicology*. Almqvist & Wiksell International. Stockholm.

Sågvall Hein, A. The LPS Inflectional Grammar. A Listing of the Rules. Department of Linguistics, Uppsala University. Forthcoming.

Sågvall Hein, A., Östling A. & Wikholm, E. 1990. Phrases in the Core Vocabulary. Center for Computational Linguistics. Uppsala University.

Sågvall Hein, A. & Sjögreen, C. 1991. Ett svenskt stamlexikon för datamaskinell morfologisk analys. En översikt. [A Swedish stem lexicon for computational morphological analysis. An overview]. In: Förhandlingar vid sammankomst för att dryfta frågor rörande svenskans beskrivning 18. Uppsala University.

Svensk Ordbok. 1986. [*A Dictionary of Swedish.*] Stockholm.

Wikholm, E. 1991. Übersetzungstheorie und maschinelle Übersetzung. Uppsala Universität. Linguistisches Institut.

Vossen, P., Meijs, W. & den Broeder, M. 1989. Meaning and structure in dictionary definitions. In: Boguraev, B. & Briscoe, T. (eds), Computational Lexicography for Natural Language Processing. Longman.

Annette Östling
Dept of Linguistics/Computational Linguistics
Uppsala University
Sweden