Developing a Core Lexicon for a Corpus-based Machine Translation System

Rebecka Edqvist
rebed@stp.ling.uu.se

Master's thesis in Computational Linguistics
Språteknologiprogrammet
(Language Engineering Programme)
Uppsala University · Department of Linguistics and Philology

14th January 2005

Supervisor:
Lars Ahrenberg, Linköping University

Abstract

This thesis concerns the development of an English-Swedish core lexicon. The core lexicon is primarily developed for the corpus-based machine translation system T4F. A bilingual core lexicon is defined as a lexicon which contains the most common words of a source language and their translations in the target language. Common words in this case are words that may appear in texts from any domain, also referred to as domain independent, or general words.

One important property of T4F, as well as many other machine translation systems, is that it is implemented into different versions for different domains. This strategy reduces ambiguity and makes better translations possible. However, it is much desirable to be able to reuse the part of an implementation that is not domain-specific. For this reason, a set of core resources are being developed for T4F, and the core lexicon constructed here will be the first part of these resources.

The core lexicon is mainly based on a parallel corpus that was compiled for this purpose. Different methods were used to extract common lemma pairs from this corpus to the lexicon. Other methods use the corpus indirectly to select words from alternative sources. Because of the varying properties of words with different part-of-speech, all word classes are dealt with separately. For most word classes, a combination of selection methods were used, e.g. selection based on distribution and frequency. One of the methods for the noun class is based on semantic criteria to find words missing from the corpus. Other sources to find words for the core lexicon are grammars, other corpora, and dictionaries.

The selected English and Swedish lemmas were given additional linguistic information about morphology, semantics, and pragmatics. This information is included in each lexical entry.

Contents

List of Tables

Acknowledgments

First of all, I would like to thank my supervisor Lars Ahrenberg for providing the topic of this thesis, and for giving me so much help and support during the work with it. I would also like to give many thanks to Maria Holmqvist, who has helped me with everything from understanding T4F to writing regular expressions. I am also grateful to all the people at NLPLAB, for helping me and for making my stay at IDA so pleasant. Finally, I want to thank Andreas, for proofreading this thesis, and for always supporting me.

1 Introduction

Machine translation (MT) is a complex and difficult task. One way to improve the result of automatic translation is to limit the efforts to a specific domain, containing texts of a similar content. This approach reduces the amount of ambiguity in the translation process, since many words have much fewer possible translations in a restricted domain. MT systems using this approach are often corpus-based, i.e. they have used translation data in the form of parallel corpora to construct their linguistic resources.

One consequence of having a system that is implemented towards a restricted domain, is that the system must be given a new set of linguistic resources: lexica, tagsets etc., every time the domain is changed. Building up these resources is a laborious process, involving compilation of a training corpus, the tagging, parsing, and linking of this corpus, and also the organisation of the extracted linguistic information.

This workload can be reduced by re-using the information which is needed for all possible domains that the system is implemented towards. This can be done by creating core resources: a set of standard solutions, resources that apply for any domain. An important part of the core resources would be a core lexicon, since much of the work during the implementation of an MT system is spent on the extraction of lexical resources.

A core lexicon would contain the most common and general words of the language. These words are domain and genre independent, i.e. they can be found in any kind of domain. In an MT environment, it would be most useful to have a bilingual core lexicon. This would in addition contain the most common translations in the target language of the selected words from the source language.

The greatest difficulty when constructing a core lexicon, is to find out which words that actually are domain independent and common. (In this thesis, the concepts of commonness and being domain independent are regarded as the same thing, although there might be a difference.) There can be no definite answer to this. The border between domain-specific and domain independent words is floating, and also depends on the semantics of words. A word may have some domain-specific sense in one domain, and another domain-specific sense in the next. For example, the word *process* will have one specific sense in computer science, another specific sense in a psycological domain, but in a domain of science, it may have one or more general senses. It should be noted, however, that a core lexicon developed for an MT system need not be a complete core lexicon for the English language as a whole. In a translation system, it is more important to have a core lexicon that covers common words of the language in domains that the system is implemented towards. Therefore, it is not a problem if the core lexicon is somewhat skewed towards such domains, as long as the constructor has a clear idea of what these domains might be.

Another issue is how the core lexicon should be constructed, and what additional information it should contain besides the actual words, to be of greatest use in the intended MT system.

A parallel corpus is a frequently used source for extraction of lexica for MT systems. If one had access to a very large and well-balanced parallel corpus for the language pair of interest, the extraction of the most common words from this corpus may perhaps be enough to make an acceptable core lexicon. However, parallel corpora are a scarce resource, especially large ones. If the parallel corpus is not large enough, it is necessary to use more criteria than only frequency to find the domain independent word in

it. A corpus may also be used as a reference for other methods to find common words from alternative sources.

1.1 Thesis Outline

In this thesis, I will explain how a bilingual core lexicon was developed for the corpus-based MT system T4F. In the background chapter, I will go through the theoretical background of this work, and introduce important concepts and methods. In the next chapter (3), I explain how the lexicon will be constructed, and what additional information it will contain. Here I also introduce the most important source of the core lexicon: a parallel corpus. This corpus was compiled as a part of this work, and is constructed to be as suitable as possible for the purpose of extracting a core lexicon.

The different word classes have different functions in the language, and this is reflected in the distribution and characteristics of the words in them. This means that different methods must be used to find words for the core lexicon for each class. The closed classes, for example, contain function words, which are frequent in all domains, and therefore they were all selected for the lexicon. For the open classes, and especially for the nouns, it was necessary to use different selection methods. One of these methods that was used for many classes, was to extract lemmas from the project corpus based on their frequency and distribution. For the nouns, a method was used that took advantage of the semantic categories of words in the corpus to find new words using WordNet.

Each word class, and the methods used to select words from it, is discussed in a separate chapter.

In Chapter 9, the result of the work will be presented. A small evaluation is presented and discussed. Here I also discuss how the lexicon can be manipulated, and how the finished core lexicon can be used in T4F.

The last chapter contains the conclusions that can be drawn from this work. Here I discuss how well the methods managed to capture the core of the language. Some suggestions of how the lexicon may be improved in the future are also given.

1.2 Aims and Objectives

The aim of this thesis is to show how a core lexicon for a corpus-based machine translation system may be developed, and which methods that may be used for this. The intention is furthermore to show how such a lexicon may be organized, and what information it should contain to be of greatest practical use. The primary application of this lexicon is the MT system T4F, and the construction of the lexicon will reflect this.

1.3 Framework

1.3.1 The KOMA project

This thesis is part of the KOMA[1]project. This is a project within the VINNOVA research program for language technology. The aim of this project is to develop methods and systems for machine translation of documents of a restricted text type (a certain domain). The project is a joint project between the NLPLAB at Linköping University and the Department of Linguistics and Philology at Uppsala University. The goals of the project are, among other things, to develop techniques and systems for lexical data generation from parallel corpora, and to improve methods and systems that have been under development at the departments with regard to their adaption to a given translation task.

[1]KOrpusbaserad MAskinöversättning = corpus-based machine translation

1.3.2 T4F

This core lexicon is being developed primarily to be used in the corpus-based MT system T4F, developed at the NLPLAB in different stages by Lars Ahrenberg, Håkan Jonsson, and Maria Holmqvist. A description of this system will be given in Section 2.3.2.

2 Background

In this chapter I will introduce concepts and methods that are central to my thesis, and also present the specific tools and techniques that I have used in relation to this. First I discuss corpus linguistics in general, and parallel corpora in particular, since a corpus forms the basis of the development of the core lexicon. I will then discuss lexicon work in natural language processing (NLP), and especially construction of bilingual lexica. Finally, I will introduce the concept of corpus-based MT and particularly the system T4F, since it is primarily in this context the core lexicon will be used.

2.1 Corpus Linguistics

A corpus is a body of text that is typically machine-readable, finite sized, and representative of a language, or some part of a language. The corpus is used for making an empirical analysis, linguistic or other, about this language. There exist different kinds of corpora for analysis of different parts of the language from various perspectives. Some examples of corpora are: spoken corpora (for investigating the spoken language), monitor corpora (used to examine the language over time), and parallel corpora (that may be used in e.g. contrastive studies). (McEnery and Wilson 2001)

In NLP, the corpus is often used to provide a system with empirical and statistical data. Typical NLP applications that often use corpora as resources are word sense disambiguation programs, parsers, and machine translation systems (Armstrong et al. 1999).

To extract information from a corpus it is necessary to go through some basic steps. First one must of course decide what kind of corpus that will be needed for the research purposes at hand. If there is not already a corpus available that fit these needs, a new one must be compiled. Texts should be sampled with a good sampling technique to be as representative as possible for the domain of interest. The corpus should also be as large as possible to reduce problems of data sparseness.

The next step is to tokenise the corpus, i.e. divide the text into lexical units. These units are most often lexical words, but may also be abbreviations, number combinations, etc.

After the tokenisation it is usual that the corpus gets annotated. Although unannotated corpora have been of much use in language study, the utility of the corpus is significantly increased by annotation (McEnery and Wilson 2001). The most basic type of annotation is to add information about metadata, e.g. title and author of the text.

2.1.1 Analysis and linguistic annotation

The most basic type of linguistic annotation is *part-of-speech tagging* (also grammatical tagging or morphosyntactic annotation). Part-of-speech tagging is accomplished by attaching a label to each word in the text, which denotes its part-of-speech. Many different tagsets exist, often containing more than the traditional word classes, and the historically most influential one is the tagset used for tagging the Brown corpus.

Lemmatisation means finding the base form for each word in the corpus, and labeling it with this information. This is an important procedure in for example lexicon extraction, since lexicon entries are

Number	Word form	Base form	Function	Morphosyntactic tag
1	she	she	subj:> 2	@SUBJ %NH PRON PERS NOM SG3
2	's	be	v-ch:> 3	@+FAUXV %AUX V PRES SG3
3	bored	bore	obj:> 6	@-FMAINV %VP EN
4	,	,		
5	he	he	subj:> 6	@SUBJ %NH PRON PERS NOM SG3
6	said	say	main:> 0	@+FMAINV %VA V PAST
7	.	.		

Table 2.1: FDG analysis of the sentence: She's bored, he said.

normally in the base form.

For some purposes it is necessary to analyse the text at a higher level of syntactic relations, i.e. to *parse* the text. Different parsers are based on different grammars, like context-free phrase structure grammars, dependency grammars, or functional grammars.

A corpus may also be semantically annotated, by labeling words with information about the semantic relationship between units in a text, or with the semantic features of words.

Annotation can be done by hand, automatically, or semi-automatically. For automatic part-of-speech taggers the accuracy is quite high: between 96% and 97% of the words may receive correct tags (Manning and Schutze 1999). Many of the automatic taggers use Hidden Markov Models, but there are also rule-based taggers like the Brill Tagger.

Syntactic parsing has been the subject of much research in language engineering, and therefore a number of automatic parsers exist.

Semantic annotation seems to be less automated, probably because of the difficulties involved in formalizing semantic information, but some semi-automatic approaches exist.

The most common format of linguistic annotation is the mark-up languages SGML (Standard Generalized Mark-up Language) or XML (a subset of SGML).

FDG

The FDG parser is the tool that was used to analyse and annotate the texts that make up the project corpus. FDG is also used to parse data in T4F.

FDG (Tapanainen and Järvinen 1997) stands for Functional Dependency Grammar, and is a syntactical analyser of sentences. The FDG parser is a tool based on the FDG technology. The output from the FDG parser contains five data fields: 1) word number, 2) word form, 3) base form, 4) functional dependency, and 5) functional tag, surface-syntactic tag, and morphological tags. See Table 2.1 for an example of FDG output.

The FDG parser is robust, so the analysis never breaks down, although the parsing can be incomplete. If no proper dependency tree could be generated for a sentence, some of the words may not get any functional tag. Also, if a word is ambiguous it can receive more than one syntactic tag. In a small evaluation in Holmqvist (2004) the FDG parser managed to give a complete syntactical parsing to between 56% and 67% of the sentences.

Problems with corpora

One of the main problems in any corpus related work is that a corpus will always to some degree be skewed and incomplete. Because it is only a sample of a population, there will be words that are missed out, and uncommon words will occur with a higher frequency than in the population as a whole. This problem is not unique for corpus linguistics, but is common for all sciences using sampling (McEnery

and Wilson 2001). The problem can be reduced by having a larger corpus and by using good sampling techniques. In Yang et al. (2002) it is shown that to automatically compile a comprehensive lexicon (i.e. containing as many words in a language as possible) one needs a corpus of such huge dimensions that present-day compiling methods would be entirely insufficient. Data sparseness, both quantitative (corpus size) and qualitative (corpus composition), is a serious drawback for most corpus-based NLP work.

2.1.2 Parallel corpora

A parallel corpus contains a source text and its translation into one or more target languages. When there is just one target language, the parallel text may be called a bitext (Ahrenberg et al. 1999).

Parallel corpora are useful, and even essential in many NLP fields, like automatic lexical acquisition (Gale and Church 1991), information extraction (Somers 1999) and machine translation. They are also much used for research in second language teaching (Botley, McEnery and Wilson 2000) and contrastive linguistics (Johansson 1997).

The overall availability is much lower for parallel corpora than for monolingual corpora. Because of this, a huge amount of research has been devoted in later years to the construction and exploitation of parallel corpora (McEnery and Wilson 2001). One fruitful approach seems to be the developing of techniques for finding parallel texts on the Web (Resnik and Smith 2003).

Work with parallel corpora goes partly through the same process as work with monolingual corpora (tokenisation, annotation etc.), but to be of practical use in research, the parallel corpora ought also to be aligned.

2.1.3 Alignment

Alignment is performed to identify which subparts of the corpus that are translations of each other. It is a mapping, or linking, between corresponding text segments (Ahrenberg et al. 1999). The alignment can be done at different segmentation levels, for example the sentence level and word level. Sentence alignment can be done automatically with a high degree of accuracy (McEnery and Wilson 2001). Word alignment is more complicated than sentence alignment, and therefore harder to automate. Some complications are that word order often differ between languages, and that boundaries between lexical units are harder to find than boundaries between sentences. Even so, there exist many automatic alignment systems based on statistical models of word correspondences, and also systems that use a combination of statistic and heuristic methods. These methods may have a success rate of almost 80% (Sågvall Hein 2004). For some purposes it may be necessary with more accurate alignment, and then one alternative may be semi-automatic systems. Such a system will be presented below.

I*Link

I*Link (Ahrenberg, Merkel and Petterstedt 2003), is the alignment tool that was used to link the parallel corpus used in this work. It is semi-automatic, and can therefore yield alignment of high accuracy.

I*Link is an interactive word aligner with a graphic interface. The interactivity means that the system gives suggestions to the user of which words or phrases that correspond and should be linked, and the user may accept or reject this suggestion. The user can also select correspondences manually.

I*Link has a set of resources that are of basically three types: static resources, dynamic resources, and patterns. Static resources do not change during a session and are typically bilingual term lists and core lexicons. The dynamic resources get updated incrementally during the linking, as words that are new to the system are added, along with new POS correspondences and syntactic functions. This learning approach helps the system to make better proposals of links in the following sentences. The pattern resources define correspondences for tokens such as cognates, numbers, and punctuation characters. (Ahrenberg et al. 2003)

I*Link also contains tools for data analysis. One is the concordance program *Link Inspector*, which can use various search criteria to find sentence-pairs of interest in the corpus. The search criteria can be combinations of word and/or base forms, POS, and word function. *Link Reporter* is a tool that can summarize and configure the information in the database, and for example produce lists of all the word-pairs in the corpus.

2.2 The Lexicon in Natural Language Processing

The lexicon is a central component of any NLP system, since it contains information about the primary component of languages, i.e words. As Grishman and Calzolari (1996) expressed it:

> "As researchers and developers in other areas of natural language processing move from toy systems to systems which process real texts over broad subject domains, larger and richer lexicons will be needed and the task of lexicon design and development will become a more central aspect of any project".

NLP applications have different goals and therefore they will also have different requirements on the lexicon. For some applications, it is necessary to have a generic full coverage lexicon, while others need a lexicon that covers the language in a certain domain. Applications that deal with only one language use monolingual lexica, while translation and second language learning systems need bi- or multilingual lexica. Applications will also vary in what over-all architecture they choose for the lexicon, and also in the formal representation of the lexical entries, and which levels of linguistic information that is kept with each entry. To quote Grishman and Calzolari (1996) again:

> "A basic lexicon will typically include information about morphology, either in a form enabling the generation of all potential word-forms associated with pertinent morphosyntactic features, or as a list of word-forms, or as a combination of the two. On the syntactic level, it will include in particular the complement structures of each word or word sense. A more complex lexicon may also include semantic information, such as a classification hierarchy and selectional patterns or case frames stated in terms of this hierarchy. For machine translation the lexicon will also have to record correspondences between lexical items in the source and target language; for speech understanding and generation it will have to include information about the pronunciation of individual words."

There are various ways to acquire a lexicon, but methods that rely on automatic or semi-automatic methods are highly preferred. Manual lexicon construction is both time consuming, costly, and error-prone.

Automatic acquisition has been facilitated by the wide availability of new electronic resources (Karagol-Ayan et al. 2003), for example machine-readable dictionaries (MRD:s), thesauri, lexical databases, and corpora.

A special kind of lexicon that is used in many NLP applications is WordNet.

WordNet

In this work, WordNet was used to find semantic categories for all the nouns and verbs in the corpus. It was also used to take out new words for the lexicon. Because of this, it deserves a special introduction.

The following definition of WordNet is given in Miller et al. (1993):

> "WordNet is an online lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, and adjectives are organized into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets."

WordNet is, as this quotation indicates, organized by semantic relations. These relations between the synonym sets, or synsets, are of different types for different word classes in WordNet. The most important relation is that of synonymy, and this is present for all categories. In WordNet, synonymy is defined in terms of substitutability. That is: two expressions are synonymous if the substitution of one for the other in a certain context does not alter the truth value.

Another relation is that of antonymy. Antonymy exists between words that have opposite senses, like *good* and *bad*. This relation is important in the organization of the adjectives and adverbs in WordNet, but it also exists for nouns and verbs.

Hyponymy/hypernymy is a semantic relation between word senses, and it is the central organizing principle of the nouns. This relation has also been called ISA relation, or subordination/superordination. The relation is defined thus: A concept represented by the synset {x, x',...} is said to be a hyponym of the concept represented by the synset {y,y',...} if native speakers of English accept sentences constructed from such frames as *An x is a (kind of) y*. (Miller et al. 1993) The hyponymy/hypernymy relation generates a hierarchical semantic structure, with more generic concepts at the top, and more specific concepts nearer the bottom. An example: *lemonade* is a kind of *fruit drink* which is a kind of *beverage*, where *beverage* is the top hypernym, and *lemonade* is the bottom hyponym.

Another important relation for nouns is the meronymy/holonymy relation, also called part-whole relation. An example is *arm* that is a holonym of (part of) a *body*.

The verbs are organized in hypernymy/troponymy relation, which is also called manner-of relation. This relation is presented more closely in Chapter 5.

2.2.1 The bilingual lexicon in machine translation

The bilingual lexicon is a critical component of any machine translation system. A significant part of the development of new MT systems is spent on constructing lexical resources (Dillinger 2001). Parallel corpora are the most important source for bilingual lexicon acquisition, but for example MRD:s, comparable corpora (Sadat, Déjean and Gaussier 2002), and OCRed dictionaries (Karagol-Ayan et al. 2003) are also used. When a parallel corpus is used as a source, the lexicon can be extracted automatically. An important part of this process is the alignment of the corpus.

Bilingual lexica, just like monolingual lexica, vary in their construction and representation depending on the needs of the NLP system. There are many different approaches to MT, and as many approaches to the lexicon construction and representation.

Some bilingual lexica are bi-directional. Bi-directionality means that the translations goes in both directions, i.e. you can look up a word in either language and get all the translations in the other language. In MT applications, the lexicon does not need to be bi-directional, since the translations goes from one source language to the target language, in most systems.

2.2.2 Core lexicon

A core lexicon is a lexicon which aims to be the "common denominator" of all possible domains in a language. In other words, the core lexicon should include the most common words: those words that may appear in any domain. A bilingual core lexicon should include the most common words of one language and their most common translations in the other language.

If a word is polysemous, the most common sense (or common senses) should be represented. This will of course only be a problem if semantic features are included in the lexicon, but for most applications this is probably the case.

The uses of a core lexicon are several. In second language learning it is useful to have access to the core vocabulary of a language for the beginner to study and learn. NLP systems that are domain-specific can use a core lexicon to start from when implementing the system towards a new domain.

The difficult part in creating a core lexicon is to find a good method for the selection of common and domain independent words. Two relevant questions are: How do one know if a word is likely to appear

in any domain, and how is commonness to be measured? The simple answer to the latter question is that commonness is the same thing as having a high frequency in a corpus. To equalize commonness with high frequency may be problematic however, because word frequency can vary a lot from corpus to corpus. If the corpus is the least bit unbalanced, some words will get unproportionally high or low frequencies. The corpus have to be well balanced and very large to correctly reflect word commonness in the language as a whole. It may therefore be wise to investigate not only the total frequency of a word, but also its distribution in the corpus. A method for this, called Distributional Consistency (DC), is proposed in Zhang, Huang and Yu (2004), and in Savický and Hlavácová (2002) three methods of "corrected frequency" are presented. The assumption underlying these methods is that if a word is common enough it will be well distributed. This also relates to the first question of this paragraph, how to know if a word may appear in any domain. If the corpus is reasonably large and well balanced, it will contain texts from many different domains, and if a word is well distributed, it will appear in many of them.

2.3 Corpus-based Machine Translation

Corpus-based machine translation includes those approaches to automatic translation that use aligned parallel corpora as a main resource when developing a translation system. Many corpus-based MT systems are restricted to translations within a certain domain. This domain could be for example a certain kind of technical texts. In a restricted domain the problems of ambiguity is often reduced considerably, and the lexicon can be much smaller and more specific.

2.3.1 Basic MT approaches

The basic approaches in MT are direct translation, transfer-based translation, statistical translation, or a combination of these (Sågvall Hein 2004).

In direct MT the most important resource is the translation lexicon. The translation is done word by word or phrase by phrase. Translation problems are handled by specific rules for each case.

In the transfer-based approach, transfer is not only done at a lexical level (i.e. translations of words and phrases) but also on a structural level. Representations of sentence structures are made by syntactical analysis of the sentences in the source language. These syntactic representation are then transferred to corresponding syntactic structures in the target language. The basic processes of a transfer-based system are analysis of source sentences, transfer of syntactic structures, and lexical items and generation (synthesis) of target sentences.

A statistical MT system is based on a model of the translation relation between two languages, and the rules of translation are acquired automatically from bilingual and monolingual corpora (Al-Onaizan et al. 1999). Most statistical MT systems use word-for-word substitution.

In all these approaches it is possible, or even necessary, to use corpora. These will be used in the implementation and training of the system. Corpora are also very useful for evaluating MT systems. The three approaches use corpora in different ways, since they are dependent on different kind of information. The direct approach can use the corpus to extract information about lexical units, for example how a particular word is translated in a certain environment. In the transfer-based approach, there will be an interest in receiving information about the structural relations. The corpus may for example be used to extract transfer rules. In the statistical approach the corpus is used to extract translation rules, and to assign probabilities to possible translations.

All approaches can use corpora to extract monolingual and bilingual lexica. Different kinds of additional linguistic and/or statistic information will be stored in the lexicon for the different approaches.

The extraction of all this information from a corpus follows the steps that were presented in the sections above, i.e. the corpus will need to be annotated, parsed, and aligned before information can be extracted.

2.3.2 T4F

The core lexicon that has been developed in this work is primarily to be used in the automatic translation system T4F.

The name T4F stands for Tokenisation, Tagging, Transfer, Transformation, and Filtering, and this is an automatic translation system that translates from English to Swedish (Holmqvist 2003). This system is based on the theory of Superlink Constraint Lexicalist Transfer Machine Translation (SCLT), as first presented in Ahrenberg (2000), and further in Jonsson (2001), and Ahrenberg and Jonsson (2001). SCLT has a direct approach to MT and therefore uses only lexical transfer and no syntactical transfer, and it localizes syntactic processing to the word level. Important features of SCLT is the concept of supertags and superlinks. A supertag is a tag encoding complex syntactic information (Ahrenberg 2000). The concept of supertags was first brought forth in Joshi and Bangalore(1999):

> "A rich description of a lexical item that impose complex constraints in a local context".

A supertag contains not only information about the word itself (part-of-speech, semantic information etc.), but also about the local context of the word. When the supertags of the source text and the target text are linked, the result is called a superlink. This gives the following definition: "A superlink is a pair of supertags, where the first element of the pair encodes the local context of a source token, and the second element of the pair encodes the local context of its translation." (Ahrenberg 2000)

The first realisation of SCLT is the prototype system SCOTS (Superlink COnstrained Transfer lexicalist translation System), presented in Jonsson (2001). In SCOTS, the supertags consist of trigram tags, plus additional relational tags (Jonsson 2001). Trigram tags are tags that hold information about the tag of the word itself (its part-of-speech, morphology, function etc.) and about the tags of the words to the left and to the right of this word. In T4F, which is a further development of SCOTS, the supertags do also contain information about dependency relations. The first implementation of T4F was developed for the ATIS domain, which is a small parallel corpus of airline travel information. The implementation is described in Holmqvist (2003) and will be shortly presented here:

The corpus data was tokenised, and then tagged with FDG. Supertags were also added to the data, which means that contextual features where added to the first tagging. The supertags are added in accordance with a set of rules.

The tagged sentences in English and Swedish were word aligned with I*Link, which was also used to create the lexical resources for the system. These resources are: monolingual lexica for English and Swedish, where words are listed with their respective tags and supertags; a transfer lexicon, containing corresponding pairs of English and Swedish word forms; a lexicon of corresponding superlinks. Some post-editing was necessary to take care of discontinuous links and null links. Sets of rules were created throughout the development of the system in order to improve its performance. These were supertagging rules, transformation rules, and filtering rules.

The implemented system was evaluated on some test data from the same domain. To translate new text from English to Swedish T4F goes through these steps:

- Tokenisation of the English text
- Tagging of the English text
- Supertagging of the English text
- Transfer the text to Swedish using the transfer lexicon, superlinks and transfer rules
- Filter alternative translations with the filtering rules
- Transform sentences (i.e. changing the word order to conform to Swedish syntax)
- Give a probability value to every possible translation.

2.4 Reuse of Linguistic Resources in MT

The reasons for wanting to reuse linguistic resources is that it can save time, labor, and money. For MT systems, that depend on large amounts of linguistic resources, this is particularly appealing. A typical reusable resource is a translated text. Reuse of translation data is the fundamental idea behind corpus-based MT. However, it is desirable to be able to reuse linguistic data not only as parallel corpora, but in a more refined form, in the shape of lexical and grammatical resources. This is because the extraction of such resources from the parallel corpora is a labor-intense process, involving steps of tagging, parsing, and aligning, as described above. The problem is that lexical and grammatical resources often are domain-specific, and their formal representation often differ between projects and systems. The solution can be to set up standards for the representation, or to make mapping algorithms between different representations (Turcato et al. 1998).

In domain-specific MT systems, the idea of reusability can be applied when a certain system is to be implemented for a new domain. It is then highly desirable to be able to reuse as much as possible of previous implementations.

2.4.1 Core resources

The point of core resources in the context of MT is to reuse resources that are independent of which domain the system is implemented towards. To these resources, domain-specific data will be added during a new implementation.

The core resources may be both grammatical and lexical resources. Also tools used to analyse linguistic data should be recycled as much as possible.

For the system T4F the core resources would consist of a core lexicon, developed in this work, and core sets of supertags and superlinks. The core set of superlinks would contain the most central grammatical correspondences at the lexical level.

3 Constructing a Core Lexicon

In this chapter I will go through the methods and resources that I have used to construct the bilingual core lexicon.

An important thing to observe is that the core lexicon developed here will not be complete. The limited size of this project, and of the material used in it, makes such a goal hard, if not impossible, to reach. Instead the ambitions are to make a foundation for the development of a more complete core lexicon. As the lexicon is used in different applications, primarily the T4F system, missing entries will be added and redundant ones removed. In this way the core lexicon will be continuously improved.

3.1 Size and Content

The size of a core lexicon can be measured in its coverage of a text from any domain. To be complete, it should cover all words that are not domain-specific, and no others. But the boundary between domain-specific, and 'general' or genre-independent words is by no means fixed, but rather a matter of degree. At one end of the scale are the word that are clearly domain and genre independent, like most of the function words (*and, to, up*). On the other end are technical terms like *dehydrator* or *dialogbox*. The problem is handling the content words in the middle of this scale. Many nouns, for example, may very well be regarded as specific for some domain, but still be common and general in the language as a whole.

An important issue when the core lexicon is bilingual, is that when the most common words in the source language are selected, their most common translations (if there are more than one) in the target language must be found. This lexicon is not intended to be bi-directional (i.e. it goes from English to Swedish, but not the other way around) so the Swedish translations does not necessarily need to be the most common words in Swedish as a whole.

The goal for the core lexicon developed here is that it should cover the common and genre independent words of an English text to an acceptable degree, and not cover those words that are clearly very domain-specific. The lexicon should also give reasonable Swedish translations to the words it covers.

3.2 Architecture

The core lexicon will be organized in separate files for every word class, for both languages. The auxiliary verbs, that are not traditionally regarded as a word class, also have their own files. Each entry in the lexicon corresponds to a line in a file, which has a unique identification number. This number connects each English entry to one or more Swedish entries, in a separate link file. The links are also stored separately for each part of speech. There is also a file that contains all corresponding lemma-pairs for each part-of-speech. These lemma pair files are not a necessary part of the lexicon, but used to simplify the process of connecting the id-numbers of the English and Swedish entries, and also give a better overview of the lexicon. Table 3.1 gives an example of how such a lemma pair file will look, and Tables 3.2 and 3.3 show an English and Swedish entry respectively.

English	POS	Swedish	POS	sem-cat	prag-cat
ability	N	möjlighet	N	cognition	tech
accident	N	tillfällighet	N	happening	
account	N	redogörelse	N	statement	

Table 3.1: An example from a lemma pair file for nouns

base	POS	morph-cat	sem-cat	prag-cat	id
ability	N	-ies	cognition	tech	ne1

Table 3.2: English noun entry

The lexicon will include certain extra-linguistic data about the lemma in each entry. This will increase its usability in the T4F system, and other systems where it may potentially be used. Every lexical entry, both English and Swedish, consists of the following: the lemma form of the word, the word's part-of-speech (syntactic category), and a unique number for identification. The entries from open classes will furthermore contain a morphological category, a semantic category, and a pragmatic category.

The **morphological category** indicates how the inflectional pattern of the word looks. The division of the categories is based on orthographical differences. For example, the Swedish verbs *köpa* and *stänga* are assigned to different categories, even though they both traditionally belong to the second conjugation. This is because *köpa* has the suffix *-te* in the past tense, and *stänga* has *-de* (a difference that comes from assimilation with the preceding consonant). In general, a typical word from every category has been used to denote it, e.g. verbs inflected in the same way as *köpa* belong to the morphological category 'köpa'.

The **semantic category** is used for word sense disambiguation, e.g. *follow* with the category 'obey' indicate the sense 'to act in accordance with some rules' (WordNet), while *follow* with the category 'travel' means 'to physically go or travel after someone or something'. The semantical categories were taken from WordNet and from grammars.

The **pragmatic category** marks from which domain the word originates, e.g. technical, fiction, or political, in this lexicon. This information is useful to have, because a polysemous word may have different senses in different domains, and thereby have different translations. E.g. *table* is translated to *tabell* in the software manuals, but to *bord* in the fictional domain. It may seem contradictory to have something like a pragmatic category in a lexicon that is supposed to contain only domain independent words. But for a word like *table*, which is relatively common and domain independent, it is most intuitive to disambiguate between the senses by means of the pragmatic context.

3.3 Compiling a Parallel corpus

A natural starting point for the construction of a bilingual core lexicon is a parallel corpus. The first part of this work was therefore concerned with the compilation of such a corpus.

The NLPLAB has a collection of parallel texts of different origin. These texts are tokenised, annotated with FDG, and sentence aligned. A corpus was collected from these texts. As explained in the background chapter, a corpus should be balanced and preferably as large as possible. This project is a small one,

base	POS	morph-cat	sem-cat	prag-cat	id
möjlighet	N	film	cognition	tech	ns1

Table 3.3: Swedish noun entry

Domain	Author and title	Origin
Technical	Microsoft Access User's Guide	LTC[1]
	Truck maintenance manuals for Scania	PLUG (Scania corpus)[2]
Fiction	Saul Bellow *To Jerusalem and Back 1976*	LTC
	Nadine Gordimer *A Guest of Honor 1970*	LTC
	J.K. Rowling *Harry Potter and the Chamber of Secrets 1998*	Sofia Helgegren
Political	Debates of the European Parliament 98-09-18	http://www.europarl.eu.int
	Debates of the European Parliament 00-01-17	http://www.europarl.eu.int

Table 3.4: The samples

however, and this is reflected in the size of the corpus. The English part contains about 56,000 tokens, and the Swedish part about 51,500 (excl. punctuations and numbers). This is very small compared to the corpora of millions and even billions of words that are compiled today.

The main reason for not making a larger corpus was that it needed to be word aligned. This was done with I*Link, a tool that is introduced in the background chapter. As is described there, this tool is interactive, and helps the user by giving suggestions of translation correspondences. Even so, the word alignment is slow work. One alternative would have been to do the linking automatically, using I*Trix, a tool that can use the resources that are built up in I*Link to do automatic translations. The performance of I*Trix was considered to be too unreliable for the purpose of this work, however. The precision of the translations is very important, since information about the frequency of lemma pairs will be used to select words for the lexicon.

The different problems that arise from using small material to make general conclusions are dealt with by methods presented further below.

3.3.1 Samples

To make the corpus balanced, it was built up by samples from different domains. The samples are taken from three domains:

1. A technical domain, consisting of texts from software manuals and truck manuals.
2. A fictional domain, with texts from novels.
3. A political domain, with protocol texts from the EU parliament.

The samples are presented more closely in Table 3.4. About 400 sentences were taken from each sample, a bit more from the truck manuals, and a bit less from the EU texts.

These sample texts show a significant variability in such features as sentence length and vocabularies. Many of the characteristics that these particular texts display, can probably be generalized to their respective domains as a whole. Some statistic data of the samples can be viewed in Appendix B.

The technical domain

Technical texts are usually characterized by short sentences. This can clearly be seen in the Scania texts, which have four to six words per sentence for the English texts, and about four for the Swedish. This is much below the average of the other texts. The Access samples does not show this pattern. Here the sentences are about the same length as the fictional texts.

[1]The Linköping Translation Corpus (Merkel 1999) contains several English-Swedish parallel texts (LTC 1999).

[2]PLUG is short for Parallel Corpora in Linköping, Uppsala, and Göteborg. URL: http://stp.ling.uu.se/ corpora/plug/. The Scania corpus is one of Uppsala´s contribution to the PLUG corpus.

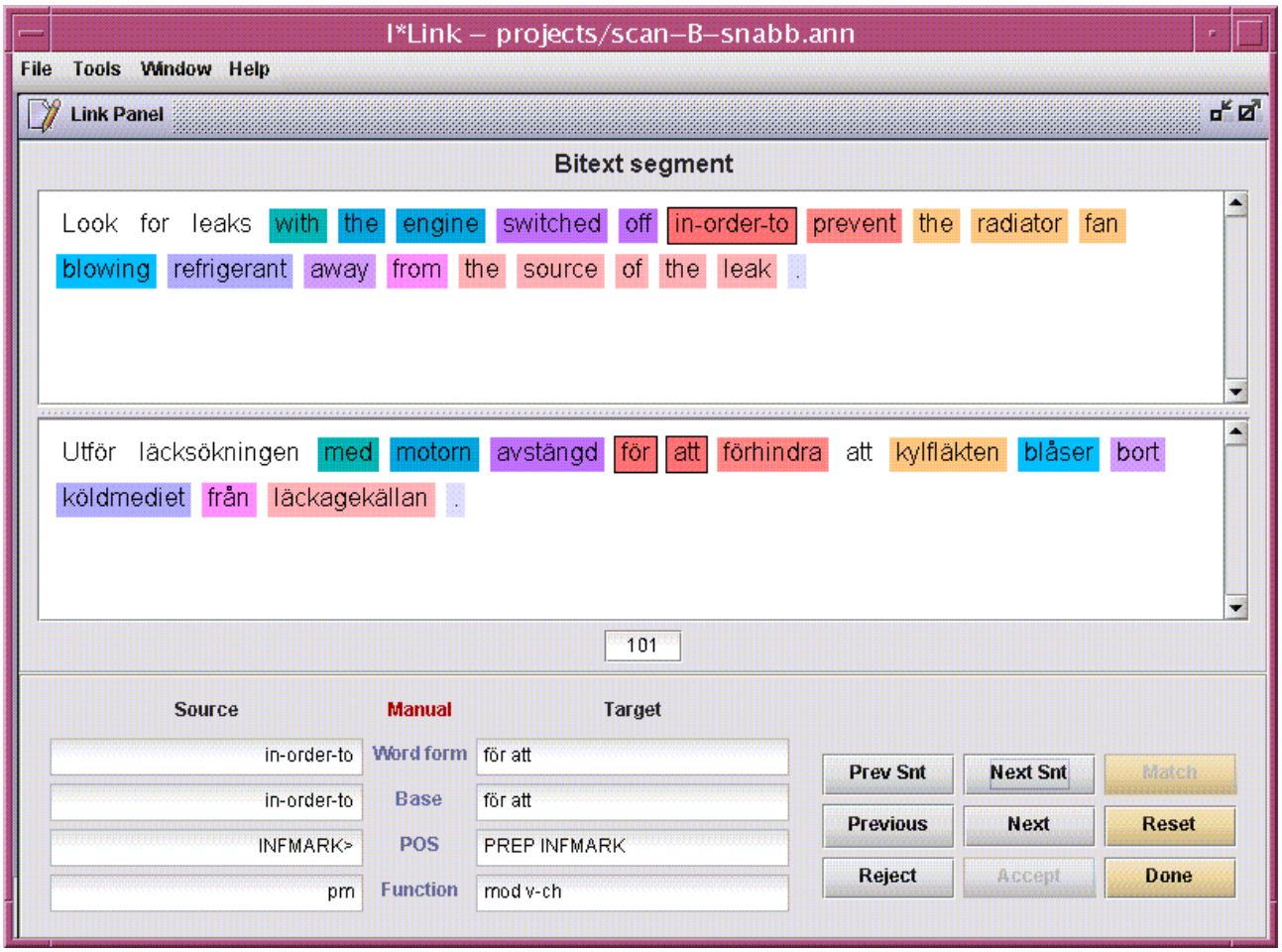

Figure 3.1: Example from I*Link

Another typical property of technical texts is that they have a restricted vocabulary. This property is apparent in the Access texts, which have the highest number of tokens per type, about eight for the English texts. The Scania texts have between four and five words, which is about the same as for the fictional texts.

The fictional domain

The fictional texts have fairly long sentences, especially the Gordimer text, which has an average of about twenty words per sentence. These texts also have a larger amount of different word types than the other domains. This is typical for this domain, which is very open, since it can contain any kind of fictional text.

The political domain

The political domain in this corpus consists of the protocols from debates in the European parliament. The sentences are long, compared to both the fictional and technical texts, with about 24 words per sentence. The token per type ratio is higher than that of the fictional texts, which means that the vocabulary is more restricted. These characteristics can probably be generalized to other political text of this kind.

Source base	Source POS	Target base	Target POS	Count (S:1313)
a	DET	en	DET	87
a	DET	ett	DET	56

Table 3.5: A 'Link report' with lemma and POS

3.3.2 Linking strategies

There is seldom only one correct solution of how to align a word or phrase in a sentence. Relatively few sentences have a one-to-one correspondence between words. Instead, so-called translation shifts are common, and the source and target sentences differ in some way. In these cases, there may be several correct ways to link the sentences. Depending on the purpose of the alignment, one of these may be preferred. A linking strategy is therefore needed to get consistent results.

The linking of the corpus in this work is not complete, since units that were deemed as irrelevant for the construction of a core lexicon have in some cases been ignored. Since I*Link gives automatic suggestions of links, these suggestion have generally been accepted when they have been correct, even if the linked pair is irrelevant for this work. If a false or no suggestion was given, irrelevant cases were left unlinked. Such irrelevant cases were: punctuations, numbers, proper names, and very rare and/or genre specific terms (often compounds). Concerning the last example, the choice may be somewhat arbitrary, but in uncertain cases the units have been linked. When the translations differ to much in structure to give a usable correspondence, the problematic words are left unlinked. An example of a problematic sentence is given in Figure 3.1, illustrating a screen-shot of I*Link. Here the phrase [look for leaks] corresponds to [utför läcksökningen] in the Swedish translation. None of the individual words in these phrases are corresponding, so the available alternative is to link the entire phrases. As this correspondence can safely be regarded as of no interest for a core lexicon, it is left unlinked. A unit that has been linked, but is with all probability useless, is [the source of the leak] that corresponds to [läckagekällan]. I*Link gave the suggestion of this correspondence, and in such a case it is easier to accept the suggestion than to actively leave the unit unlinked. It was not necessary to pay any attention to whether the word forms corresponded, only lemma forms are of interest for this work.

Some of the sample material (one Access text and the Harry Potter text) had already been linked in other projects. These links could be reused, but the samples were carefully investigated so that the linking strategy used in them did not deviate in any essential way from the one presented here. The sample from Harry Potter was word aligned by Sofia Helgegren, and the Access text by Magnus Merkel.

3.3.3 Lemma lists

From the word aligned corpus, lists of unique lemma pairs with their respective part-of-speech tags were extracted. This was done with the tool Link Reporter. Each lemma pair had its frequency included. The Table 3.5 shows how the output from Link Reporter looks.

Some basic editing was performed on these lists, like removal of numbers and punctuations. The lists from the samples where then concatenated into one list of all the lemma pairs in the corpus. This list contained just over 11,500 unique pairs.

3.4 Extracting the Lexicon

3.4.1 Strategies

Many of the most frequent words in any corpus will be words from the closed classes. These are the so called function words: the conjunctions, prepositions, pronouns, determiners, and certain adverbs. These make up almost a third of the project corpus. Since the closed classes contain words that are both common and of a finite quantity, it was decided that all words from the closed classes should be included in the lexicon. This was a rather straight-forward process, which will be presented below.

The main focus of this work regards the matter of which words to include from the open classes: nouns, verbs, adjectives, and adverbs.

As mentioned earlier, the corpus used in this work is quite small. The problems of skewedness and data sparseness that this leads to could clearly be seen in the material. The skewedness means that some words get a high frequency because they are common in a particular text. This is particularly evident in the class of nouns. The five most common nouns in the corpus were: *field*, *data*, *commission*, *table*, and *vehicle*. These nouns come from only one domain each, the word *commission* from the political domain, all the others from the technical. These words do not intuitively seem like very common words in English as a whole, a feeling that can be confirmed by examining their position in the Brown corpus. Only the word *field* is placed among the hundred most common nouns there, and it is likely that the restricted sense of this word in the software manual texts where it appears, is not in so many cases equal to the senses that this word has in Brown. The Brown corpus will be introduced further down.

Concerning the problem of data sparseness, this might be harder than the skewedness to spot at a first look. But if words that are not so common occur with a high frequency in the corpus, one may conclude that more common words will get unproportionally low frequencies. A comparison with the Brown corpus shows that of its hundred most common nouns about five are not in this corpus at all, and several of the common nouns appear only once.

Obviously, a strategy is needed to handle the skewedness and data sparseness of the corpus. Since all the open classes have different behavior they were dealt with separately.

3.4.2 The Brown corpus

The Brown corpus was used as a comparison and a complement to the project corpus. For the open classes, words that are among the five hundred most common in Brown, and not already selected by other methods, were added to the lexicon.

The original name of the Brown corpus is: The Standard Corpus of Present-Day American English. The corpus consists of just over 1 million words of running text of edited English prose, printed in the United States in 1961. (Francis and Kucera 1964) The corpus is divided into 500 samples from 15 different domains. It is available in both a tagged and an untagged version. The tags hold information about part-of-speech, and some basic morphological information. Some words have their own tags, like *not*, and existential *there*.

3.5 Producing Equivalent Pairs of Wordforms

The morphological categories can be used to produce all possible forms of a word. In the application of the lexicon in a machine translation environment, it would be useful to be able to link each wordform in the source language to its equivalent wordform in the target language.

I will present a suggestion of how this could be done for this core lexicon. I will use nouns to exemplify, but the principle would be the same for all the classes that have morphological categories.

First, one would need some rules to produce all wordforms for each individual word. In the case of English nouns, the morphological category ’-ies’ looks like this:

cat-name	stem	s-nom	pl-nom	s-gen	pl-gen
’-ies’	-(C)y	+y	+ies	+’s	+ies’

The (C) that is inserted in front of the y in the ’stem’ column indicates that a consonant must immediately proceed the y for nouns of this category. Words like *boy* belong to the category ’-s’ instead.

A rule that produced all possible wordforms would say something like: remove the letter under ’stem’ from the end of the lemma, add the contents in each form column, e.g. s-nom (singular nominative) or pl-gen (plural genitive), to the stem to produce that form. This rule would produce the wordforms *city, cities, city’s, city’s*. For the irregular nouns there would be no need for rules to produce the wordforms, since all wordforms are already listed for each of them.

The Swedish morphological category ’stad’ looks like this:

cat-name	stem	pl-stem	s-nom	s-nom-def	s-gen	s-gen-def	pl-nom	pl-nom-def	pl-gen	pl-gen-def
stad	-	städ	+	+en	+s	+ens	+er	+erna	+ers	+ernas

A similar rule as the one for English nouns, but one that also handled the plural stem, would produce: *stad, staden, stads, stadens, städer, städerna, städers, städernas*. Then another rule would be needed that could state that the English form ’s-nom’ is equivalent to the Swedish forms ’s-nom’ and ’s-nom-def’, and that ’s-gen’ is equivalent to ’s-gen’ and ’s-gen-def’, and so on. This could gives pairs of the form (city; stad, staden), (city’s; stads, stadens), (cities; städer, städerna) and (cities’; städers, städernas).

Some nouns might need special rules to be handled correctly. The noun *clothes*, for example, which can not have a singular form. At present, this word is marked with a ’0’ for all singular forms, and it would be necessary to have a rule that took care of this.

4 Nouns

The noun class has the largest number of words of all classes. It is a considerable challenge to decide which words from this large and diverse class that qualify for a core lexicon.

This corpus contains 16,124 English nouns, which constitute 27% of the entire corpus. As could be seen in the previous chapter, the nouns show a marked imbalance towards domain-specific words. Noticeable is also that even though most of the very common nouns (according to the Brown corpus) do appear in the corpus, many of them appear only once or twice. This may be an indicator that other common nouns are not in the corpus at all. A solution to this problem will be proposed below.

4.1 Using Semantic Categories to Select Nouns

The main idea is to use the semantic categories of the words that are already in the corpus, to find other words that are as common and general as these. A semantic category consists of a set of words (and concepts) that are related in meaning, like 'body parts', or 'mammals'. The assumption is that if a certain amount of words from the same semantic category are present in the corpus, then other words from that category should also be candidates for the core lexicon. An example: the words *Monday*, *Tuesday*, *Wednesday*, and *Thursday* all occur in the project corpus. These words can all be assigned to a semantic category called 'days of the week'. The fact that *Friday*, *Saturday*, and *Sunday* are not in the corpus must be regarded as incidental, and these words will also qualify as candidates for the lexicon. The three domains contain texts of different semantic character, and therefore bring words from different semantic fields into the lexicon.

That words belong to the same semantic category does not always mean that they have the same commonness, as they do in the example of weekdays. Consider the semantic category 'body parts': It may contain both the very common noun *hand* (24th most common noun in Brown), and the quite uncommon word *earlobe*. This shows that frequency must also be taken into account, when using this method.

To give semantic categories to the nouns, WordNet was used.

4.2 Nouns in WordNet

As was explained in the background chapter, the nouns in WordNet are hierarchically organized in hypernymy/hyponymy relations. The hierarchy can be seen as a lexical inheritance system where a hyponym inherits the distinguishing features of its hypernym, and also has a feature of its own, that distinguishes it from the hypernym and from other hyponyms at the same level. (Miller et al. 1993) An example: *common canary* inherits the feature of being a "small Old World finch" from its hypernym *canary*, but it also has its own distinguishing feature of being a "yellow cage bird noted for its song". *Canary* in its turn has inherited the feature of having a "short stout bills adapted for crushing seeds" from its hypernym *finch*. As can be seen, the hypernymic relation goes from generic to specific.

The nouns are separated into a number of hierarchies with some generic concept at the top. Examples of such “top” concepts are *entity*, *abstraction*, and *psychological feature*. The hierarchies correspond to relatively distinct semantic fields.

The semantic category that is given to each word from the corpus, is that word’s immediate hypernym in WordNet.

4.3 Selecting Nouns for the Core Lexicon

A semantic category from WordNet was assigned to all noun-pairs that occurred more than once in the corpus. The noun-pairs with a single occurrence were removed, since they were considered as too uncertain to use as a basis for the method. The removal of the single occurrences reduced the amount of unique noun-pairs to less than half the size.

As explained above, the theory is that if a certain amount of words from the same semantic category are present in the corpus, then other words from this category are also qualified for the lexicon. The amount of words that were needed from a certain category were set to three. These three words were automatically qualified for the lexicon. The selection of new nouns proceeded as follows: A noun is given a semantic category, that is this noun’s hypernym in WordNet, e.g. *bathroom* is given the semantic category ’room’. The members of a semantic category are its immediate hyponyms, for ’room’ these are for example *classroom*, *dining room*, and *kitchen*. If three different members from one category appear in the corpus, then all its members are considered as candidates for the lexicon. In the example of ’room’, there were more than three hyponyms present, so this category was used to select new nouns. The Brown corpus is used as a reference to remove candidates that are too uncommon. In the case of the category ’room, the word *manor hall* was excluded, among others, but for example *library* was included in the lexicon.

The nouns that were selected from WordNet in this way were given Swedish translations before they could be inserted in the lexicon. These translations come from Norstedts English-Swedish dictionary, and they were selected with consideration of the semantic category of the English word.

In addition to the nouns from the semantical categories, the most common and well-distributed (= appearing in more than one domain) nouns from the project corpus were included in the lexicon. The most common nouns from the Brown corpus were also included. Some of the English nouns from these sources were already in the lexicon, thanks to the semantic category criterium. In these cases, it was decided whether this noun did have some more common sense, and/or a more common translation. This decision could be made by consulting WordNet and the English-Swedish dictionary (c.f. Stora engelsk-svenska ordboken (1989)). If this was the case, this noun was added again in another entry, with the more common semantic category and translation.

4.4 Morphological Categories

Each noun in the lexicon is assigned to a morphological category based on its inflectional pattern. English nouns have only four different forms (nominative case singular and plural, and genitive singular and plural). They have in this work been divided into four different regular categories, and a few irregular ones. The regular categories are called ’-s’, ’-es’, and ’-ies’, and they cover a large majority of the nouns in the lexicon. The irregular nouns have their own categories, e.g. ’man’ and ’knife’. In the case of the English nouns, the list of irregular nouns include only those words that appear in the lexicon.

The Swedish nouns have eight possible inflections for every noun, and these forms show a considerable diversity in spelling between different nouns. A list of morphological patterns for Swedish nouns was borrowed from the morpological descriptions in Sve.Ucp (Sågvall Hein 1998), where every category is represented by a pattern word. Some new categories were added to the list, and some categories were removed. It contains 66 different patterns, and is intended to cover most Swedish nouns.

English	POS	morph-cat	sem-cat	prag-cat	id
account	N	-s	statement		ne4
accuracy	N	-ies	quality		ne5
action	N	-s	act	pol	ne6
Swedish	**POS**	**morph-cat**	**sem-cat**	**prag-cat**	**id**
aktivitet	N	film	process	fic	ns4
anordning	N	stol	instrumentality		ns5
ansikte	N	samhälle	external body part	fic	ns6

Table 4.1: Samples from the English and Swedish lexicon files for nouns.

Appendix A shows all the morphological categories.

4.5 Lexicon Sample

A sample of how the English and the Swedish lexicon files for nouns look is given in Table 4.1. The nouns that lack pragmatic category come from either WordNet or the Brown corpus.

5 Verbs

The verb class is much smaller than the noun class, and verbs are also more polysemous than nouns (Miller et al. 1993). The most common verbs show the most polysemy (*have, be, run, go* etc.). The semantic interpretation of these is heavily dependent on the nouns with which they co-occur. These facts are confirmed in the project corpus when looking at the most common verb, *be*. This was given 23 different translations in the corpus. Consequently, the challenge when it comes to the verbs in a bilingual core lexicon, is to include the most general meaning, and thereby also the most general translation.

The verb frequencies in the corpus seem to be less skewed than the nouns. A comparison to the Brown corpus shows that the most frequent verbs show a similar distribution in both corpora. Some of the verbs that are very frequent in the project corpus are clearly genre-specific though, like *create* and *add* (used in the senses of creating and adding data-base objects in the MS Access samples).

Because verbs are very polysemous, it is quite desirable that they have semantic categories. WordNet was used to find semantic categories for the verbs from the corpus. Intuitively, it is not as easy to form distinct groups based on semantic similarities with verbs as it is with nouns. WordNet was investigated to see if new verbs could be extracted by using semantic categories, with the same method as for the nouns.

5.1 Verbs in WordNet

Verbs in WordNet are arranged primarily by a relation called troponymy. If a verb V1 is a troponym of another verb V2 (thereby the hypernym of V1), then V1 specifies a certain manner of carrying out V2 (Fellbaum 1998). For example, *wash down*, *slurp*, *gobble*, and *devour*, are all troponyms of the verb *eat*. Troponymy builds hierarchical structures similar to the hyponymy relation for nouns. There are differences, however: “verb trees are flatter than the noun trees and rather more like bushes, rarely exceeding four levels” (Fellbaum 1998). Because each tree hierachy is smaller, containing fewer of the total number of verbs, there exist more different hierarchies for verbs than for nouns.

Verbs are also related by antonymy (see Section 2.2), and by entailment. An example of the latter: *eating* entails *swallowing*, that is, when someone eats it is necessary that s/he swallows. Entailment goes in one direction, since it is not true that *swallowing* entails *eating* (i.e. one can swallow without eating).

5.2 Selecting Verbs for the Core Lexicon

The verbs were given semantic categories by looking up their hypernyms in WordNet. The question is if these hypernyms can be used to find verbs that should be part of a core lexicon, as was done with the nouns. One problem with this approach is, that whereas the nouns in WordNet are arranged in a relatively small number of deep hierarchies, the verbs are arranged in very many shallow ones. Many verbs are therefore at the top of a hierarchy. This is of course an effect of the differing properties of the nouns’ hyponymy relation compared to the verbs’ troponymy relation (and this in its turn is a result of the differing properties of the noun and the verb class). An examination of the troponymy relation in WordNet shows that it is less appropriate for finding new verbs, using the same method as for nouns.

cat-name	non-3-sing	3-sing	ing	past
walk	walk	walk+s	walk+ing	walk+ed

Table 5.1: A morphological category of English verbs: ’walk’

cat-name	base	pres	pret	sup	perf-part
stänga	stäng	stäng+er	stäng+de	stäng+t	stäng+d

Table 5.2: A morphological category of Swedish verbs: ’stänga’

An example can illustrates this: The verb *help*, in the sense ’give help or assistance; be of service’ has the hypernym *support*. The verb *support* has besides *help* the troponyms *sponsor, patronize, promote, undergird,* and *second*. It is not intuitive to say that if *help* is in the corpus, then other ’manners of supporting’ should also be candidates for the core lexicon.

Since the verb class is smaller and the verb distribution in the project corpus is less skewed than the noun class, the need of for a ’selection technique’ is less acute. Instead the verbs were selected for the lexicon based on distribution in the project corpus. The Brown corpus was used to fill up the lexicon with its most common verbs (those not already included from the project corpus). The auxiliary verbs were regarded as function words, and dealt with separately.

5.3 Morphological Categories

The inflectional systems of verbs in English and Swedish verbs are quite similar. They can both be given a set of regular categories into which the majority of the verbs can be sorted. An example of a regular category in English is ’walk’, which is shown in Table 5.1. Here it can be seen that there is one column for each form. E.g. the form *walked*, in the column called ’past’, is used for both the past tense and the past participle, and the base form *walk* is used both for present tense (non 3:rd singular) and the infinitive. An example of a Swedish regular form is given in Table 5.2.

Both English and Swedish also have a limited set of irregular verbs that include many common verbs. For example *come, came, came*, and *go, went, gone* in English, and in Swedish *komma, kom, kommit* and *gå, gick, gått*.

For this lexicon, the majority of the irregular verbs in English and Swedish have been listed with their own morphological categories, also those that do not appear in the lexicon. There are 161 English irregular categories listed, and 180 Swedish. Only a sample of these are included in Appendix A.

5.4 Lexicon Sample

A sample of how the English and the Swedish lexicon files for verbs look is given in Table 5.3.

English	**POS**	**morph-cat**	**sem-cat**	**prag-cat**	**id**
add	V	walk	increase	tech	ve10
address	V	pass	communicate/intercommunicate	fic	ve11
adjust	V	walk	change/alter	tech	ve12
Swedish	**POS**	**morph-cat**	**sem-cat**	**prag-cat**	**id**
ange	V	ge	tell	fic	vs8
anpassa	V	ropa	change/alter	tech	vs9
anse	V	se	judge	pol	vs10

Table 5.3: Samples from the English and Swedish lexicon files for verbs

6 Adjectives

Adjectives are, just like verbs, highly polysemous and frequently get their meaning determined by the noun that they modify (Fellbaum 1998).

An informal comparison between the adjectives of a high frequency in the project corpus, and those in the Brown corpus, shows that their distribution is quite similar. Of the thirty most common adjectives in Brown, all but one (*public*) were also more or less common in the project corpus. The adjectives showed some skewedness towards domain-specific words. For example, *dangerous* is the forth most common adjective in the project corpus, but not particularly common in Brown. Other examples of adjectives that are much more common in the project corpus are *related* and *financial*. These adjectives all appear in only one domain.

6.1 Adjectives in WordNet

It is desirable to assign semantic categories also to the adjectives. They can not be taken from WordNet, as with the nouns and verbs, however.

WordNet does not organize adjectives in hierarchical structures, like verbs and nouns. This is because "Nothing like the hyponymic relation that generates nominal hierarchies is available for adjectives: it is not clear what it would mean to say that one adjective 'is a kind of' some other adjective." (Miller et al. 1993) They are instead organized in clusters centered around two antonymic adjectives (Fellbaum 1998). The antonymic pair in the center are called direct antonyms. Direct antonyms occur with great frequency in the language, and they appear together in word association tests. They "constitute a conspicuous but small part of the adjective lexicon" (Fellbaum 1998). The rest of the adjectives are classified as semantically similar to one or the other of the direct antonym pair. For example, *astronomic*, *enormous*, and *extensive* are all semantically similar to *large*. They are considered to be 'indirect antonyms' of *large's* direct antonym *small*.

6.2 Selecting Adjectives for the Core Lexicon

As shown in the previous section, adjectives can not be given semantic categories from WordNet. It may be possible to find some other source of semantic categories for adjectives, but this was judged to be outside the scope of this thesis. Instead, adjectives were mainly selected from the project corpus, based on their distribution and frequency.

The organizing relation for adjectives in WordNet, the antonymy relation, could be used to find potentially missing words. This is because the so-called direct antonyms are common and central to the language (Fellbaum 1998), and if one half of a pair is present in the corpus, then the other half should also be a candidate. A closer inspection of the antonymic pairs in WordNet reveals that some antonyms are less than common, however. This is particularly the case when the antonym is constructed with one of the prefixes un-, in-, or non-. Some examples from the corpus are, *real-unreal*, *mechanical-nonmechanical*,

cat-name	base	comp	super
hard	hard	hard+er	hard+est
careful	careful	more careful	most careful

Table 6.1: Two morphological categories of English adjectives: ’hard’ and ’careful’

cat-name	utr	neutr	plur	best-mask	komp	super	sup-best
kall	kall	kall+t	kall+a	kall+e	kall+are	kall+ast	kall+aste
kall(m)	kall	kall+t	kall+a	kall+e	mer kall	mest kall	mest kalla

Table 6.2: Two morphological categories of Swedish adjectives: ’kall’ and ’kall(m)’

ready-unready. In these pairs the first word is considerably more common than the second (checked in the Brown corpus). Consequently, frequency need also be taken into consideration.

Frequent adjectives from the Brown corpus were also included in the lexicon.

6.3 Morphological Categories

In English, adjectives are only inflected for comparison. The comparison can be done in two different ways: by adding a suffix, or by using ’more’ and ’most’. The regular forms fall into five different categories, including comparison with ’more’ and ’most. The most common of the regular categories, ’hard’, is shown in Table 6.1. The category using ’more’ and ’most’ is also shown.

Other categories are ’big’, where the final consonant is duplicated before the suffix (*bigger*, *biggest*) and ’wide’, in which are included words that end with an ’e’, and because of this are suffixed with ’r’ and ’st’ (*wide-r*, *wide-st*).

A list of irregular categories has been created, but it is not exhaustive. It contains for example *good*, *better*, *best*.

The Swedish adjectives have in addition to the compared forms, different inflections for gender and number of the noun that they modify. The list of categories is intended to cover all adjectives, but some irregular form might be missing. An example of a morphological category for Swedish is ’kall’, exemplified in Table 6.2 The Swedish adjectives can, in similarity to the English adjectives, be compared either with suffixes, as in the example above, or with ’mer’ and ’mest’. Since the Swedish adjectives are also inflected for gender, there cannot be only one category for the ’mer’, ’mest’ form of comparison, as in English. In those cases where an adjective that is compared with ’mer’, ’mest’ is inflected for gender in the same way as in a category compared with suffixes, like ’kall’, this adjective is given a new category named as the other category with a (m) added, e.g. kall(m). This is also shown in Table 6.2. An adjective of this category e.g. *mekanisk* would consequently have the forms: *mekaniskt*, *mekaniska*, *mekaniske*, *mer mekanisk*, *mest mekanisk*, *mest mekaniska*.

6.4 Lexicon Sample

A sample of how the English and the Swedish lexicon files for adjectives look is given in Table 6.3.

English	**POS**	**morph-cat**	**prag-cat**	**id**
basic	A	careful	tech	adje15
big	A	big	fic	adje16
black	A	hard	fic	adje17
Swedish	**POS**	**morph-cat**	**prag-cat**	**id**
bra	A	bra	tech	adjs18
demokratisk	A	kall(m)	fic	adjs19
direkt	A	kort(m)	pol	adjs20

Table 6.3: Samples from the English and Swedish lexicon files for adjectives

7 Adverbs

The adverb class consists of both function words and content words. New adverbs are primarily derived from adjectives by adding the suffix -ly. The goal for the core lexicon is that it should contain all the functional adverbs, and the most common of the content adverbs. WordNet was not used for selecting adverbs, nor for giving them semantic categories. The adverbs in WordNet are organized in antonymic relations, just like the adjectives, so no semantic categories could be extracted. The adverbs ending with '-ly' have information about root adjective. An inspection of the adverbs in WordNet showed that many of them did not have any antonym listed, and therefore the antonymic relation was not used to find new adverbs, as was done with the adjectives.

To find adverbs for the lexicon, the project corpus was used, but also some English grammar books. These grammars, Longman English Grammar (Alexander 1988) and Collins Cobuild English Grammar (John Sinclair et. al 1990), where used to take out common adverbs, and also to give these semantic categories. The project corpus was used to extract frequent adverbs for the lexicon. Lastly, the Brown corpus was used to insert common adverbs.

One problem related to using adverbs from the corpus is that words of this class, especially those that have a more functional role, often are difficult to classify correctly. This is a problem that is shared by all functional classes. Many function words are homonyms, i.e. they have the same form, but belong to different word classes. For example, the English word *that* can be either a pronoun, an adverb, an adjective, or a conjunction (Oxford Reference Online 1999). The reason that this is problematic for this work, is that the FDG parser often tags these words incorrectly. This means that some words from the corpus are tagged as adverbs when they should not, and vice versa. This happens for nouns, verbs, and adjectives also, but not as often, and these mistakes are easier to spot since e.g. the lemma form is strange. The list of adverbs for the lexicon was looked through for obvious mistakes, but in doubtful cases the FDG tags were kept.

An additional method that might have been used to give more adverbs to the lexicon, is to add the -ly suffix to the selected adjectives. This could be done on the presumption that the common adjectives would become common adverbs. A cursory inspection shows that this is true in some cases, for example *certain* gives *certainly*, and *actual* gives *actually* which are both among the more common adverbs, according to the Brown corpus. However, the methods presented above for selecting adverbs produced enough adverbs for the lexicon, if the number was compared to the number of words from the other classes in the lexicon. It was assumed that the most common adverbs would be selected with these methods. However, if the lexicon need to be enlarged in the future, or if common adverbs do seem to be missing, the method suggested here could very well be used.

As mentioned, semantic categories for the adverbs were taken from grammars. Many adverbs clearly falls into categories like time, place, and manner etc. The adverbs of time in the lexicon are all of the type 'indefinite time', e.g. *again*, as opposed to adverbs of definite time, like *last week* or *next Sunday*, which are often whole phrases not appropriate for the core lexicon. Other categories are focus adverbs (*only*), adverbs of indefinite frequency (*often*), and adverbs of degree (*rather*).

English	POS	morph-cat	sem-cat	prag-cat	id
after	ADV		place		adve9
afterwards	ADV		indef-time		adve10
again	ADV		indef-time	fic	adve11
Swedish	**POS**	**morph-cat**	**sem-cat**	**prag-cat**	**id**
automatiskt	ADV		manner	tech	advs12
baklänges	ADV		place		advs13
bakom	ADV		place		advs14

Table 7.1: Samples from the English and Swedish lexicon files of adverbs

7.1 Morphological Categories

Some adverbs can be compared in the same way as adjectives, both in English and Swedish, but most adverbs are indeclinable. An example of an adverb that can be inflected is in English *fast*, *faster*, *fastest*, and in Swedish *snabbt*, *snabbare*, *snabbast*. Adverbs of this type have been marked with appropriate morphological categories in the lexicon, in the same way as the other open classes.

7.2 Lexicon Sample

An example of how the English and the Swedish lexicon files for adverbs look is given in Table 7.1. Many of the adverbs are taken from grammars, and therefore have no pragmatic category. None of these adverbs can be inflected, and therefore have no morphological categories.

8 The Closed Classes

As mentioned in Chapter 3, all members of the closed classes: determiners, conjunctions, prepositions, and pronouns, will be part of the core lexicon. Also the auxiliary verbs are considered to be function words, and are all included in the lexicon.

There is therefore no need of a method of selection (though some selectional choices need to be made here too, as shall be seen) but these classes have other difficulties that need to be dealt with. One of these difficulties is that many of the function words can belong to more than one word class. In these cases, it is important that this word is given all its possible parts-of-speech. Sometimes a word may be quite uncommon in one word class, and should only be given the more common part-of-speech.

To find all English function words, grammar books (Longman English Grammar (Alexander 1988) and Collins Cobuild English Grammar (John Sinclair et. al 1990)) were primarily used. An English-Swedish dictionary (Stora engelsk-svenska ordboken 1989) was used to translate the result. The project corpus, and in some cases also the Brown corpus, was used as a complement to the grammars.

The words from the closed classes are not given any semantic, morphological, or pragmatic categories.

8.1 Pronouns

Most of the pronouns for the lexicon could be found in the grammars. They are marked with categories that tell what kind of pronoun it is, and in relevant cases, of person, case, number, and gender, e.g. *she* is marked 'pers-nom-sg3-f'.

Somewhat problematic are the indefinite pronouns. This is a large and diverse group, and different sources give different answers to which words that belong to it. In general, this lexicon will follow the tagging given by FDG in these cases. For example, the words *many* and *much* are tagged as either determiners, adverbs, or pronouns by FDG, while according to Norstedts English-Swedish dictionary, these words can be only adjectives or adverbs. In the core lexicon the FDG parser's tagging is followed, and these words are given the parts-of-speech suggested by it. The project corpus was thus used as a complement to the grammars to find the more problematic pronouns.

8.1.1 Lexicon sample

An example of how the lexicon files for the pronouns look is given in Table 8.1.

8.2 Prepositions

To find the prepositions, Collins Cobuilds Prepositions (John Sinclair et. al 1992), was used. However, the Brown corpus was consulted to investigate the frequency of the prepositions from this book, and the most unusual ones were removed. Finding appropriate translations to the prepositions was not an easy task. The most common prepositions are very polysemous, and therefore have many possible translations.

English	POS	morph-cat	id
fewer	PRON	indef	pne14
he	PRON	pers-nom-sg3-m	pne15
her	PRON	pers-obj-sg3-f	pne16
Swedish	**POS**	**morph-cat**	**id**
bägge	PRON	indef	pns11
de	PRON	pers-nom-pl3	pns12
de där	PRON	dem-pl	pns13

Table 8.1: Samples from the English and Swedish lexicon files for pronouns

English	POS	id
about	PREP	ppe1
above	PREP	ppe2
according to	PREP	ppe3
Swedish	**POS**	**id**
av	PREP	pps1
bakom	PREP	pps2
beträffande	PREP	pps3

Table 8.2: Samples from the English and Swedish lexicon files for prepositions

It is quite likely that in a large enough corpus, all common prepositions in English could be found as translated to any Swedish preposition.

For example, *of* is conventionally translated as *av*, but in the project corpus it is given 17 different translations including *av*, e.g. *mellan*, *i*, and *på*.

This problem was dealt with by checking which translations the prepositions had received in the project corpus, and include those translations that where used more than once. This gave 14 translations of *of*. The most frequent and conventional translations, according to the corpus and also Norstedts dictionary, were marked as such. In the case of *of*, the most common translation was *av*. If the word was not represented in the corpus, or had received a translation that was judged as too uncommon, a translation from the dictionary was used instead.

8.2.1 Lexicon sample

An example of how the lexicon files for the prepositions look is given in Table 8.2.

8.3 Conjunctions

The conjunctions (and subjunctions, which were dealt with at the same time) were mainly taken from the grammars. But also in this case, the project corpus was used as a complement. The translations to the conjunctions and subjunctions were looked up in Norstedts English-Swedish dictionary when needed, and in some cases the project corpus was consulted to see which translations were most common. The lexicon files for the conjunctions and subjunctions look just like those for the prepositions.

English	POS	id
another	DET	dete6
any	DET	dete7
both	DET	dete8
Swedish	**POS**	**id**
alla	DET	dets2
allt	DET	dets3
andra	DET	dets4

Table 8.3: Samples from the English and Swedish lexicon files for determiners

8.4 Determiners

A determiner is “a member of a mainly closed class of words that precede nouns (or, strictly speaking, noun phrase heads) and limit the meaning in some way.” (Chalker and Weiner 1998). The class of determiners seem to be more established in English grammar than in Swedish. In fact, neither the Swedish grammar consulted for this work, Thorell (1977), nor Norstedts English-Swedish dictionary, use the syntactic category determiner. Words that are counted as determiners in other sources are here listed as adverbs, adjectives, or pronouns. The Brown corpus do not use the category in the same way as the English grammars. Instead it has categories called for example ’pre-qualifier’ and ’post-determiner’, that together correspond to to the class of determiners. However, the FDG parser uses determiners, and these have been included in the lexicon, together with determiners from the English grammar books. Most of the determiners in the lexicon are also represented with other parts-of-speech. Possessive pronouns (*my*, *her*) are listed as determiners in the English grammars, and therefore given both categories in the core lexicon, to mention one example. The FDG parser is a bit inconsistent in that it sometimes tags corresponding words as determiners in English (e.g. *last* in ’last night’), but as adjectives in Swedish (*förra* in ’förra natten’), in an equivalent context. This inconsistency is kept in the lexicon, since FDG is going to be the primary parser in the future applications (T4F) of the lexicon.

8.4.1 Lexicon sample

An example of how the lexicon files for the determiners look is given in Table 8.3.

8.5 Auxiliaries

The English auxiliaries were taken from the grammars, and their translations were taken from the dictionary. The translations are not always classed as auxiliaries in Swedish. The word *may*, which is translated to *få*, is an example of this.

9 Results

In this chapter, I will present the finished lexicon that is the result of this work (Section 9.1), and also a small evaluation (Section 9.2). In Section 9.3, I will discuss how this lexicon can be used in T4F.

9.1 The Finished Core Lexicon

The architecture of the core lexicon was outlined in Chapter 3. This architecture could mainly be followed, with the exception that adjectives did not receive semantic categories. Another natural deviation is that those words that were taken from other sources than the project corpus do not have a pragmatic category.

The resulting lexicon contains 2177 English entries and 2252 Swedish entries, making up 2603 unique lemma pairs. The size of the lexicon is primarily decided by the size of the project corpus. The reason that the number of lemma pairs in higher than the number of English and Swedish entries is explained by the fact that one entry may have more than one translation. Table 9.1 shows some statistical facts about the lexicon.

Some comments about the number of words in each category might be needed: The determiners, and also the prepositions, have a lot more Swedish entries than English. In the case of the determiners, the main reason for this is that some of the determiners are inflected for gender and number in Swedish. Each inflected form have been given a unique entry. An example is *which*, that is given the translations *vilken*, *vilket* and *vilka*. In the case of the prepositions, each of the common English prepositions were given some alternative translations from the project corpus, at the most fourteen different translations (the preposition *of*). These alternative translations are to some extent the same for many of the English prepositions, e.g. the Swedish preposition *på* appears fifteen times. But apparently, the variation is greater on the Swedish side.

One important thing to note here is that the number of unique lemmas is quite a bit smaller (about half) than the total number of entries implies, since there is one entry for every semantic and pragmatic category that a word has. For example, the verb *ask* has eight entries, since it has three different semantic categories, each having at least two pragmatic categories (Table 9.2). The reason for having an entry for each of the values of these categories, is that they may all have different translations. In the case of *ask*, the semantic category ’communicate, intercommunicate’ gives the translation *fråga*, the category ’demand’ gives the translation *uppmana*, and *ask* with the semantic category ’request, bespeak’ is translated to *be*. In this example, it is only the semantic categories that give different translations, but the pragmatic categories may very well do so too, especially if there are no semantic categories, as in the case of the adjectives.

It would be possible to have only individual entries for each unique lemma, and let this have the semantic and pragmatic categories as attributes. The English and Swedish entries would then need to be matched based on the values of these attributes, as well as the identification number. In the present version of the lexicon, however, the method used was considered as the easiest to implement and to use practically, and with no obvious drawbacks.

	Noun	Verb	Adj	Adv	Det	Pron	Prep	Conj	Aux	TOT
Eng	745	553	316	300	40	88	82	40	13	2177
Swe	761	575	304	300	70	114	72	42	14	2252

Table 9.1: Number of entries in each class in the lexicon

base	**POS**	**morph-cat**	**sem-cat**	**prag-cat**	**id**
ask	V	walk	communicate/intercommunicate	fic	ve27
ask	V	walk	communicate/intercommunicate	pol	ve28
ask	V	walk	communicate/intercommunicate	tech	ve29
ask	V	walk	demand	fic	ve30
ask	V	walk	demand	pol	ve31
ask	V	walk	demand	fic	ve32
ask	V	walk	request/bespeak	fic	ve33
ask	V	walk	request/bespeak	pol	ve34

Table 9.2: Lexicon entry for 'ask'

9.2 Small Evaluation

In this small evaluation, the nouns of the lexicon were compared to the nouns of a tagged version of the LOB corpus (Johansson 1986). The LOB (Lancaster-Oslo/Bergen) corpus is the British counterpart of the Brown corpus, consisting of one million words of British English texts from 1961.

The best way to evaluate the core lexicon would have been to compare the entire lexicon to some parallel corpora. However, there are practically no aligned English-Swedish parallel corpora available that may be used for this purpose. It was supposed that a comparison of only the English side would also give valuable information. The reason for choosing to compare only the nouns instead of all classes, is that the lemma forms from the lexicon could then be used directly. Since the LOB corpus is not lemmatised, it would have been necessary to produce all (or at least some) of the wordforms for e.g. the verbs in the lexicon, to compare them. Although information about wordforms is present for each entry, there are no implemented rules for producing these. Another reason for choosing to compare only the nouns, is that it would probably be less informative to compare the other classes to a monolingual corpus. The nouns are less polysemous than e.g. verbs and adjectives, and therefore have fewer possible translations. For other classes than nouns, especially for the closed classes, it is probably more important to examine if they have acceptable translations. Furthermore, the noun class is the largest and most diverse of the word classes, and therefore the most difficult to select lexical entries from. For these reasons, it was assumed that a comparison of only the nouns would give some hint of how well the lexicon covers a balanced corpus.

The comparison was done only with the nouns in singular nominative form (tagged 'NN') in LOB, since this form is equal to the lemma form of the nouns from the lexicon. The result showed that the nouns of the lexicon covers almost exactly a third of the nouns in the LOB corpus (55,297 out of 206,115). This number may not be so informative in itself. More interesting is to learn if those nouns that were covered, were indeed general and common, and if those not covered were more domain-specific.

The nouns that were not covered by the lexicon were summarily examined, to see if a pattern could be discerned. This examination was done in a selection of the individual domains, or categories, of the LOB corpus. There are fifteen categories, and picked out for examination were the following: 'Press: reportage', 'Religion', 'Learned and scientific writings', and 'Adventure and western fiction'. The most frequent of the LOB nouns that were not present in the lexicon were compared with the totally most frequent nouns in a category. The assumption was that if the lexicon covers the general and genre-independent

Press: reportage		Religion		Scientific writing		Adventure & Western	
missing	total	missing	total	missing	total	missing	total
meeting	year	book	life		time	sir	man
party	time	gospel	**book**	group	case	gun	time
labour	council	text	word	blood	number	mouth	door
chairman	**meeting**	verse	man	surface	part	arm	way
company	committee	prayer	church	method	school	wall	voice
defense	government	question	**gospel**	influence	field	train	face
conference	man	spirit	day	equation	area	window	hand
market	week	earth	**text**	scale	point	mother	room
group	**party**	politics	work	temperature	air	hair	night
ministry	way	proof	verse	view	water	father	head
industry	night	saying	time	trade	system	coach	side
commonwealth	war	soul	**prayer**	length	work	pocket	girl
building	car	peace	**question**	sodium	value	party	house
Russians	work	knowledge	world	test	use	judge	car
game	**labour**	chapter	mind	force	form	blow	day
son	day	view	place	desire	effect	lot	life
correspondent	**chairman**	religion	fact	range	solution	knife	boy
attack	air	conference	**spirit**	behaviour	fact	glass	**sir**
society	**company**	baptism	way	analysis	theory	doorway	road
league	world	seed	power	resistance	level	corner	moment
dividend	part	ministry	death	party	**group**	card	thing
county	court	principle	grace	movement	way	boat	**gun**
trade	money	sin	body	knowledge	rate	wind	**mouth**
officer	**defense**	congregation	use	education	example	soldier	**arm**
capital	**conference**	worship	law	unit	age	horse	name
association	town	blessing	**earth**	technique	table	ground	mind
training	**market**	argument	**politics**	coal	model	engine	light
education	office	pleasure	**proof**	pressure	**blood**	cigarette	back
visit	side	catechism	**saying**	question	section	hat	water
peace	number	version	name	importance	period	drink	morning

Table 9.3: Nouns from some categories in LOB: the most common missing nouns from the lexicon, and the totally most common in the category with missing nouns marked in bold.

words, then the list of uncovered words should display a much higher amount of genre-specific words.

Table 9.3 shows the 30 most frequent nouns from four of the categories, together with the 30 most frequent words that were not covered by the lexicon, presented in frequency order. The nouns that are not covered by the lexicon are marked in bold style in the column of the totally most common nouns in the category. It is apparent that a lot of the common words have been removed, like *time, life*, and *man*. In the list that is best covered, scientific writing, only two words of the thirty most frequent ones are not covered. This is quite significant, since this category is the largest in LOB (120,000 words and 27,000 singular nouns), containing many different kinds of texts. This makes it more likely that the frequent words of this domain are of a general and common kind.

The lists of uncovered words apparently display a higher amount of genre-specific words. But words that are not necessarily as closely related only to this genres do occur among the uncovered words, for example *son* (press), *book* (religion), *group* (science), and *mother* (adventure). At least one word that might be viewed as genre-specific have instead been covered by the lexicon, the noun *committee*. That this particular word is covered is most certainly a consequence of the fact that a political domain was used to build the lexicon.

As it has been discussed earlier in the thesis, it is not a simple matter to divide between words that are domain-specific and domain independent. The noun *blood*, for example, that is very frequent in the scientific category, also appeared in the religion category, and in the adventure and western fiction category. In these categories *blood* might very well be regarded as domain-specific too, with a different semantic interpretation in each domain. However, with the definition used in this thesis, if a word occur frequently in many different domains this should be motivation enough to regard it as domain independent.

An examination was also performed to see if there were any nouns from the core lexicon that were not present in the LOB corpus at all. There were almost thirty such nouns. However, a closer investigation proved that some of these word, e.g. *today*, *home*, and all month and weekday names, where tagged as a special kind of noun (NR), closer related to adverbs, in LOB. The rest of the words are such that they have different spelling in British English and American English. These were *humor, labor*, *odor*, *program*, and *center*. LOB contains only British English, where these words are spelled *humour, labour*, *odour*, *programme*, and *centre*. This is very likely the reason that these words were left out. The core lexicon is not intended to have a preference for either British or American English, but the fact that the Brown corpus is one of the sources for it has apparently left a trace.

It might be risky to draw any conclusions from this evaluation regarding how well the rest of the lexicon covers the core of English. The nouns are partly selected using other methods than the rest of the classes. But on the other hand, the other open classes: verbs, adjectives, and adverbs, are all smaller than the noun class, and it could clearly be seen in the project corpus that these classes contained less domain-specific words, showed more similarities between the domains and also with the Brown corpus. These things considered, it is likely that the other open classes would show at least as good results as the nouns, in a corresponding comparison. When it comes to the closed classes, these are all part of the lexicon already, so a comparison with a monolingual corpus would be quite uninformative.

9.3 Manipulating the Lexicon

The small evaluation showed that the lexicon do cover a lot of the common words in the tested domains. It also showed that some common words were not covered, and this is of course a problem. However, this result was partly foreseen, because of the difficulties involved in finding the domain independent words of the language, and because of the limited size of this project. The lexicon developed here should not be seen as a complete and closed set of entries, but instead as a foundation for a more complete core lexicon to which entries should be added and removed. This is what will happen when the lexicon is used in T4F, where new entries will be added as the system gets implemented towards new domains, and common words that appear in this domain proves to be missing from the system. In this aspect, it is more important that the lexicon is constructed in such a way that it is easy to change.

Below I present some guidelines to how the lexicon can be altered by inserting, deleting, and updating entries.

Insertion

For an entry to be added, the following needs to be done:

1. Change the English and Swedish words that are to be inserted into their respective lemma forms, and identify their part-of-speech. This needs not be the same for the English and the Swedish word.
2. Find the lemmas' respective morphological categories in the list that has been made in this work. For some words with irregular inflection, a new category must be created.
3. Look up the lemmas' semantic categories in WordNet. (This is always the same for both the English and Swedish word).
4. Give the words a pragmatic category. If the words come from a domain that is already represented in the lexicon, this category can be used, otherwise a new one can be created.
5. Add this information to the lemmas in a way that corresponds to the existing entries in the lexicon files.

The words should then be entered as a pair in the correct word pair file, and also individually in the English and Swedish lexicon files, according to their part-of-speech. The monolingual lexicon files are arranged in alphabetic order, so the words should be entered in consideration of this. The words will then need to be given identification numbers. This number could either be the next consecutive number of the last entry in each file, or if the alphabetic order is to be kept, the number of the word that follows the newly entered word in the file. In the first case, the id:s of the new English and new Swedish entries should then be connected and given a link id in the link-file. In the latter case, all words following the new one must be given new numbers, and the link-file with the corresponding id-numbers must be updated with the new numbers.

If the English or Swedish word is already present in the lexicon, with the exact same additional information, only the new corresponding word need to be added in its lexicon file, and also the new pair and the new id link.

Deletion

For an entry to be deleted, these step should be followed:

1. Find the English and the Swedish words in the monolingual lexicon files, and remove them.
2. Remove their id link from the link file.
3. Remove the pair from the word pair file.
4. Update the id numbers in the monolingual lexicon files.
5. Update the id links in the link file.

Update

If an attribute of some entry, for example the morphological category of an English word, needs to be changed, this can be done by simply finding the word that is to be changed, delete the old value and enter the new one. This will not affect any other part of the entry, or any other part of the lexicon. If it is the semantic or pragmatic category that is to be changed, this must be done in both the English and the Swedish monolingual lexicon files, as well as the word-pair file.

If the future applications of the core lexicon turn out to require a large amount of changes in the lexicon, it might be preferable if the lexicon is stored in a database, which is more suited for manipulating data.

9.4 The Application of the Lexicon

The primary application of this lexicon is the corpus-based machine translation system T4F, presented in the background chapter. I will here discuss some ways in which the core lexicon may be used in this system.

First, it might be a good idea to give the words in the lexicon supertags that are consistent with those used by T4F. The additional linguistic information that is stored with each entry in the lexicon can be regarded as tags of the words. To map these tags into supertags, contextual information will be added. For the nouns, this could for example be information about whether the English word is preceded by a determiner (e.g. 'the'), or not. This information decides how the word is translated, since Swedish has different forms for the definite and the indefinite form. Recall the example from Section 3.5, about how to produce pairs of equivalent wordforms. If the nouns are given supertags, the redundancy of letting the form *city* be paired with both *stad* and *staden* can be removed. In this case, *city* with the tag 'pos:n - case:nom - num:sg - def:**indef**' will be paired with *stad*, and *city* tagged 'pos:n - case:nom - num:sg - def:**def**' will be paired with *staden*. Table 9.4 shows the supertags of all wordforms of *city* and *stad*.

It is the intention that T4F in the future should have a set of core resources, that will be used when the system is implemented towards a new domain. During such an implementation, resources for the new domain are extracted from a training corpus. Among these resources are monolingual lexica for the source and target languages, and also a transfer lexicon. Since a training corpus is a sample of the texts in a domain, it may be missing words that are present in some other part of the domain, that later is translated by the system. The core lexicon will therefore be used to complement the domain-specific lexica with general and common words, that may appear in one text from the domain, but not another.

Another way in which the core lexicon can be used, is to help FDG to parse the training corpus. The parsing is done before the alignment, to improve this. In the English version of the FDG parser, it is possible to add words to the internal lexicon that is used by FDG to parse a text (Holmqvist 2003). The information that is stored with each entry in the lexicon can be seen as the tags of these words. These tags contain more information than the tags given by the FDG parser, namely semantic and pragmatic information. If the parser has access to this information in its internal lexicon, the words for which this apply can be given this semantic and pragmatic information in their tags, in addition to the information FDG normally gives. In this way, all words in the training corpus that are represented in the core lexicon, are given semantic and pragmatic information. The lexicon may also be used to give tags to words that are not present in it, but display a similar character. If a word from the training corpus has certain features that are equal to the features of a word from the core lexicon, then other features of the lexicon word may be applicable for the new word. For example, if two word have the same semantic category and part-of-speech, it is likely that they will behave similarly and appear in the same context.

Finally, the core lexicon may be used in the alignment phase of the implementation. The alignment tool I*Link uses a bilingual lexicon as a resource to find corresponding units in the source and target languages. The core lexicon may be used to expand and improve this lexicon.

English:	supertag
city	pos:n-case:nom-num:sg-def:indef
city	pos:n-case:nom-num:sg-def:def
city’s	pos:n-case:gen-num:sg-def:indef
city’s	pos:n-case:gen-num:sg-def:def
cities	pos:n-case:nom-num:pl-def:indef
cities	pos:n-case:nom-num:pl-def:def
cities’	pos:n-case:gen-num:pl-def:indef
cities’	pos:n-case:gen-num:pl-def:def
Swedish:	supertag
stad	pos:n-case:nom-num:sg-gend:utr-def:indef
stads	pos:n-case:gen-num:sg-gend:utr-def:indef
staden	pos:n-case:nom-num:sg-gend:utr-def:def
stadens	pos:n-case:gen-num:sg-gend:utr-def:def
städer	pos:n-case:nom-num:pl-gend:utr-def:indef
städers	pos:n-case:gen-num:pl-gend:utr-def:indef
städerna	pos:n-case:nom-num:pl-gend:utr-def:def
städernas	pos:n-case:gen-num:pl-gend:utr-def:def

Table 9.4: Nouns with supertags

10 Conclusions

In this chapter, I will discuss the results of this work, and the methods I have used. I will also discuss how the lexicon may be improved.

Some important questions need to be discussed: Do the lexicon that is the result of this work contain the “core” of English, and its most common translations? Did the methods used to select words produce a satisfactory result? Is this result better than other possible method would have produced, like for example a simple extraction of the most frequent words from a corpus?

I will try to answer these questions in the following sections.

10.1 The Core of the Language

What is the core of the language? This is of course an important question to answer before there can be any definite answer as to whether or not the lexicon covers it. And the answer will be different depending on your reasons for asking it. The reasons in this thesis are that we want to find those words that can be part of any restricted domain, so that these words can be part of core resources of T4F. Other applications that may also be interested in the core of the language, for example in the area of language learning, may not care so much if some more domain-specific words are part of a core lexicon. In a translation system, however, this may lead to more ambiguity, if a word has one sense in one domain, and another in the next. Another matter is that the core that is of interest for this lexicon, is probably not really a core of the “whole” language, however this may be defined. More interesting is of course the core of all domains that the T4F system may be implemented towards. This is mainly texts of a technical or administrative character.

To conclude this discussion, one can say that there are many words that are borderline cases. In the application of T4F, the future will prove which words are really needed in the core lexicon, and which are not. In any case, there are some words that are very common in most domains, and the evaluation showed that many of these are indeed covered by the lexicon.

Of all the nouns in LOB, thirty percent were covered by the lexicon. And as argued above, the nouns are probably the most problematic class, so there is no reason to believe that other classes would show a worse result.

10.2 The Methods

The other questions to consider is how well the methods used to select words for the lexicon worked, and if they produced a better result than other possible methods.

It may be necessary to recall which methods that were actually used in this work, to be able to answer this question.

Here follows a short summary: A parallel corpus was compiled out of ten parsed, tagged, and sentence aligned bitexts from three different domains. This corpus was aligned at the word level. The corpus was then used to select entries for the core lexicon. The entries were selected with different methods for

each part-of-speech. In the process of selecting words from the open classes, the words were also given additional linguistic information, in the form of morphological, semantic, and pragmatic categories. The methods to select words from the corpus were mainly based on distribution and frequency. In the case of the nouns, the semantic categories of the words in the corpus were used to select nouns that were not represented in the corpus. The selected entries were separated into two corresponding monolingual lexica, with unique identification numbers that linked them together.

10.2.1 The project corpus

One important method to consider is thus how the compilation of the project corpus was done. This thoroughly affects the construction and content of the lexicon.

An alternative method that would seem very appealing, is to select a core lexicon by simply collecting the most frequent words from an already existing parallel corpus. It is possible that in a large enough (several million words) and well balanced corpus, the most frequent words would also be the most common and general in the language as a whole. It has been discussed before in this thesis that even a very large corpus will have some degree of skew, but this skew affects uncommon words more than common. In a small corpus, however, the problems of skewedness and data sparseness will be seen for all words, and this makes it necessary to use other methods to find the common words. To compile a sufficiently large corpus for this simple selection method to be used, would be a project far exceeding the frames of a thesis work.

Another possible method would be to extract the most common words from a monolingual English corpus. There do exist very large corpora of this kind, e.g. the British National Corpus, which consist of 100 million words. If such a method was to be used, the Swedish translations would need to be taken from other sources, like dictionaries. This would be unsatisfactory, since words often have many different translations. A problem would then be to choose the right translation depending on the sense of the word. Should for example *table* be translated into *bord* or *tabell*? Even when you have the semantic category of the English word, the same information may not be present for the alternative translations in a dictionary. You would also lose information about whether a word have different translations for different domains, i.e. the pragmatic information. Finally, in a dictionary there is no consistent information about which translation is the most common for some word, and for the core lexicon this can be important information.

A project corpus was instead compiled, and since it would not be very large, it was important that it was as well-balanced as possible. Therefore samples from three different domains: technical, fictional, and political, were used. Each of these domains contributed with both general and also domain-specific words to the corpus. The reason for choosing these particular domains is is that technical and political, or administrative, domains are the kind of domains that the T4F system is most likely to be implemented towards in the future. If the core lexicon was intended to contain the core of the language as a whole, it would of course not be desirable that it had a skew towards certain domains. In this context, however, it is not a problem if the lexicon contains a larger amount of technical and administrative words, than words of a less relevant character, as long as these words are still relatively common.

It is less likely that the system will be implemented for translation of fictional texts, since this domain is very open and large, and thereby difficult to translate automatically. Because of these characteristics, however, this domain may contribute with more general words to the corpus than the other domains can.

Samples were taken from different texts of the domains, to increase the balance of the corpus. In the case of the technical domain, this actually consists of two sub-domains, one of software manuals, and the other of truck maintenance manuals.

Despite the efforts to make the corpus well balanced, the limited size of it made the problems of data sparseness and skewedness quite apparent. The methods used to select words for the lexicon are all aimed at minimizing the negative effect of this.

10.2.2 Selecting lexicon entries

The methods that were used to select verbs, adjectives, and to some extent adverbs and nouns from the project corpus, are based on the frequency and distribution of the lemma pairs in the corpus. Those pairs from each class that were sufficiently well-distributed and frequent were selected for the lexicon. This method seems to have produced a satisfactory result, although this can not be certified since no proper evaluation was done with the verbs, adjectives, and adverbs.

In the case of the adjectives, the words that were selected from the corpus were complemented by antonyms from WordNet. This is motivated by the fact that the antonymic pairs common and central to the language (Fellbaum 1998). However, a frequency criterium was also applied, to avoid uncommon words like *unready*.

10.2.3 Selecting nouns from WordNet

The noun class is (as has been concluded before in this thesis) the most problematic class to select common words from. The nouns in the corpus show a large degree of skew, and therefore the method used for the other classes to select word do not suffice. If only nouns from the project corpus were to be selected, this would result in both many unwanted (i.e. domain-specific) entries, and also many missing entries in the core lexicon.

The primary method to find common nouns not present in the project corpus, was based on the fact that nouns can quite easily be divided into semantic categories. The assumption is that nouns that belong to the same category are approximately as general and common, so if words from a certain category are present in the corpus, then other words from this category will automatically be candidates for the lexicon. WordNet was used to assign such semantic categories to the nouns from the corpus.

One thing that became apparent as semantic categories were assigned to the nouns, is that WordNet is probably not ideal for this purpose. WordNet is very extensive, and aims to cover all words in the language, and also all the different senses that a word may have. It was sometimes hard to choose the right sense for an individual word, and in some cases several senses in WordNet covered the sense of a word in the corpus.

An example of this is the word *country*, with five different senses. In the following example, the immediate hypernym of each sense (which will be the semantic category of the word in the lexicon) is also included.

1. **country, state, land** – (the territory occupied by a nation; ”he returned to the land of his birth”; ”he visited several European countries”)

   → **administrative district, administrative division, territorial division** – (a district defined for administrative purposes)

2. **state, nation, country, land, commonwealth, res publica, body politic** – (a politically organized body of people under a single government; ”the state has elected a new president”; ”African nations”; ”students who had come to the nation’s capitol”; ”the country’s largest manufacturer”; ”an industrialized land”)

   → **political unit** – (a unit with political responsibilities)

3. **nation, land, country, a people** – (the people who live in a nation or country; ”a statement that sums up the nation’s mood”; ”the news was announced to the nation”; ”the whole country worshiped him”)

   → **people** – ((plural) any group of human beings (men or women or children) collectively; ”old people”; ”there were at least 200 people in the audience”)

4. **country, rural area** – (an area outside of cities and towns; ”his poetry celebrated the slower pace of life in the country”)

   → **geographical area, geographic area, geographical region, geographic region** – (a demarcated area of the Earth)

5. **area, country** – (a particular geographical region of indefinite boundary (usually serving some special purpose or distinguished by its people or culture or geography); ”it was a mountainous area”; ”Bible country”)
   → **region** – (a large indefinite location on the surface of the Earth; ”penguins inhabit the polar regions”)

Even though it is possible to understand these different senses, it can be difficult to decide which one that applies to the word, when appearing in a new context (probably more so when English is not your first language). This is a sentence taken from one of the EU texts:

> “Madam President, I am extremely surprised at Mr Kerr’s reaction because he should clearly recall the day when we agreed, and subsequently voted, in the committee to remove references to *countries*.“

From this isolated sentence, it is possible to eliminate senses 3 and 4, at once, and also quite possibly sense 5. Left are senses 1 and 2, which in my opinion are quite similar. If more context is taken into consideration, the correct sense might perhaps be chosen, but in this case, both categories were assigned to the word. This choice was supported by the fact that the word ’country’ appeared at several places in the corpus, where it a bit more clearly belonged to either sense 1 or sense 2. Since the assignment of semantic category is done to the word type ’country’, and not the individual tokens, all the examples of the word are considered when the semantic category is chosen. In the example of *country*, the problem of selecting sense is not so great, since it appear many times in the corpus.

When two, or maybe three senses have all been as appropriate for some word in the corpus, they have all been assigned to the word type. This may result in a slight redundancy in the lexicon, if there are many entries for the same word that are alike except for the semantic category. But if only one sense was randomly chosen, this would affect the result of the selection method, since new words are selected based on how many words that are in the same category, and different senses give different categories. The number of senses for *country* is normal or even low, at least for the more common nouns in WordNet. Up to twenty-nine different senses (the word *line*) occur.

Another problem with WordNet for this work, is that it has such a fine-grained hierarchy. This can sometimes make semantically similar words, which would probably be found under the same category in a less detailed system, appear under different categories. An example of this is ’hand’ and ’foot’. These words are closely related semantically, and because of this the presence of one in the corpus, should with the semantic method make the other one a candidate for the lexicon. In WordNet, these words are not under the same hypernym: ’hand’ is found under the semantic category ’extremity’, and ’foot’ under ’vertebrate foot’, which in its turn is a hyponym of ’extremity’.

A solution to this may be to somehow look at several levels, and not only the immediate hypernym. This is more difficult to motivate and to practically handle, though. One could, for example, include information about the second hypernym of a word, what could be called its grand-mother (in tree notation), for each noun. Then this information could be used in a similar manner as the information about the immediate hypernym, so that if the same word appears as second hypernym for some number of nouns, all the nouns that are hyponyms two levels down, ’grand-children’ of this word, will be candidates for the lexicon. This method was implemented as a possible complement to the method using only the immediate hypernym. It was not used for the lexicon at this point, since it seemed to be too unreliable, and the result contained a lot of material that was unusable for the lexicon.

This method would in any case not help in connecting the words ’hand’ and ’foot’, since these words are in an ’aunt - niece’ relationship. There may be methods to exploit this relation also, but the risk is that it would give a lot of unwanted information. An example of all existing hyponyms (all ‘descendants’), of the word ’extremity’, included in Appendix C, may illustrate this. The number of arrows before a word shows how deep in the hierarchy it is. The immediate hyponyms of ’extremity’ (there are nine) thus have one arrow before them, and they are, together with all their descendants, separated by extra newlines from the next immediate hyponym. It can be seen in this example that a lot of words are irrelevant for the content of a core lexicon. These words are very technical and uncommon in the language as a whole, like

'dactyl' and 'nipper'. But a few are much more common, like 'finger' and 'arm'. Another thing that can be seen in this small sample, is that the position in the hierarchy does not decide whether a word is more common in the language. The organization of WordNet only concerns setting more general concepts higher than more specific ones, and general concepts need not be denoted by general words.

Because of these difficulties, only the immediate hypernym-hyponym relation was used to select nouns from WordNet. It is possible that this has lead to the result that common words such as 'foot' are missed out, since there is little chance that categories such as 'vertebrate foot' can be given to other words in the lexicon (see Appendix C, for the immediate hyponyms of 'vertebrate foot'). To be selected for the lexicon the category must be given to at least three different words.

When a category from WordNet has qualified for the lexicon, the words in it are compared to the Brown corpus. If they are frequent and well-distributed enought, they are selected for the lexicon. However, some categories proved to be quite inappropriate for the selection of new nouns. The motivation for this method was that words that appear in the same semantic category should be related and approximately as common, so that if one appeared in a context, another may likely appear in a similar context. Some of the categories, e.g. the category 'artifact', does not fit this description. The words in the corpus that have this category are only three: *surface, thing,* and *goods*. Words that would be selected from this category with this method are for example *block, cone, decoration, textile* and *opening*. These words do not seem to be very closely related semantically, even if they all may be called 'artifacts'. To say that if the word *surface* (with the sense "the outer boundary of an artifact") is in the lexicon, the word *cone*, or *textile* should also be there, seems quite unintuitive. The solution to this problem has been to remove categories that display this kind of 'unrelatedness' from the categories that are candidates for the lexicon. Fortunately, most of the categories could be kept.

An alternative, in view of these problems, would have been not to use WordNet at all, and instead use some other semantic hierarchy. WordNet has the big advantage of being freely and easily available, however. And also, one of the things that made the use of WordNet problematic in this work, that it is so detailed and exhaustive, could be an advantage when assigning semantic categories to words in some small domain for T4F. A less exhaustive hierarchy would contain fewer of the technical and domain-specific words of this domain.

10.3 Further Development and Possible Extensions

One extension of the core lexicon that will be done before it is used in T4F, is to implement rules to produce pairs of corresponding word forms from the corresponding lemmas, in the way that was described in Section 3.5. In a translation system, it is very useful to have direct access to corresponding word forms, and not only corresponding lemmas. The core lexicon will also be given supertags, as described in Section 9.4.

As has been mentioned earlier in this thesis, the core lexicon developed here is intended to be a foundation for a more complete lexicon. However, changes may not only be needed for the content of the lexicon, by insertions and deletions of entries. Since the lexicon has not been practically used in T4F when this thesis was written, it is possible that some other aspect of the lexicon is not so well suited for this system. It is the intention, though, that all aspects of this lexicon and its architecture and construction should be so well described in this thesis, that changes can be done without great difficulty.

A possible extension for this lexicon in the future, would be to add probabilities to the entries. This would further help applications like T4F to choose the right translation of an ambiguous word. This would be very helpful e.g. for the common prepositions, which have many possible translations.

Bibliography

Ahrenberg, L. (2000). Superlinks: A new approach to constraining transfer in machine translation, PLUG Project Report.

Ahrenberg, L. and Jonsson, H. (2001). From Word Alignment to Machine Translation via Superlinks, *Proceedings of the 13th Conference on Computational Linguistics*, Uppsala.

Ahrenberg, L., Merkel, M. and Petterstedt, M. (2003). Interactive Word Alignment for Language Engineering, *The 11th Conference of the European Chapter of the Association for Computational Linguistics April*, Agro Hotel, Budapest, Hungary, pp. 12–17. project note.

Ahrenberg, L., Merkel, M., Ridings, D., Sågvall Hein, A. and Tiedemann, J. (1999). Automatic processing of parallel corpora: A Swedish perspective, at Linköping University Electronic Press, Computer and Information Science Series.

Al-Onaizan, Y., Curin, J., Jahr, M., Knight, K., Lafferty, J., Melamed, I., Och, F., Purdy, D., Smith, N. and Yarowsky, D. (1999). Statistical machine translation. Final Report, JHU Workshop 1999.

Alexander, L. G. (1988). *Longman English Grammar*, Longman Group UK Limited.

Armstrong, S., Church, K. W., Isabelle, P., Manzi, S., Tzoukermann, E. and Yarowsky, D. (eds) (1999). *Natural Language Processing Using Very Large Corpora*, TEXT, SPEECH AND LANGUAGE TECHNOLOGY, Volume 11, Kluwer Academic Publishers, Dordrecht.

Botley, S. P., McEnery, A. M. and Wilson, A. (2000). *Multilingual Corpora in Teaching and Research*, Rodopi, Amsterdam.

Chalker, S. and Weiner, E. (1998). ”determiner”, *The Oxford Dictionary of English Grammar*, Oxford Reference Online, Oxford University Press. 14 October 2004.

Dillinger, M. (2001). Dictionary Development Workflow for MT: Design and Management, *MT Summit VIII*, Santiago de Compostela, Spain.

Fellbaum, C. (1998). A Semantic Network of English: The Mother of All WordNets, *Computers and the Humanities* **32**: 209–220.

Francis, W. N. and Kucera, H. (1964). *Brown Corpus Manual*, Brown University, Providence, Rhode Island. Revised 1971, Revised and Amplified 1979.

Gale, W. A. and Church, K. W. (1991). Identifying Word Correspondences in Parallel Texts, *in* M. Kaufmann (ed.), *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, Asilomar, California, pp. 152–157.

Grishman, R. and Calzolari, N. (1996). *Survey of the State of the Art in Human Language Technology*, chapter 12: Language Resources. 12.4 Lexicons.

Holmqvist, M. (2003). Översättningssystemet T4F - en implementation för ATIS-domänen, Unpublished report.

Holmqvist, M. (2004). *Identifying Translation Shifts Using Dependency Grammar and Interactive Word Alignment*, Master's thesis, Linköping University.

Johansson, S. (1986). *The Tagged LOB Corpus Users' Manual*, Norwegian Computing Centre for the Humanities, Bergen.

Johansson, S. (1997). *Practical Applications in Language Corpora*, Lodz University, chapter Using the English-Norwegian Parallel Corpus - a corpus for contrastive analysis and translation studies, pp. 282–296.

John Sinclair et. al (ed.) (1990). *Collins Cobuild English Grammar*, HarperCollins Publishers.

John Sinclair et. al (ed.) (1992). *Collins Cobuild English Guides, 1 Prepositions*, HarperCollins Publishers.

Jonsson, H. L. O. (2001). *Exploring Superlink Constrained Lexicalist Transfer for Empirical Machine Translation*, Master's thesis, Linköping University.

Karagol-Ayan, B., Doermann, D. and Dorr, B. (2003). Acquistion of Bilingual MT Lexicons from OCRed Dichtionaries, *Technical report*, Institute for Advanced Computer Studies (UMIACS), University of Maryland.

Manning, C. D. and Schutze, H. (1999). *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, Massachusetts.

McEnery, T. and Wilson, A. (2001). *Corpus Linguistics*, Edinburgh University Press. Second edition.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K. (1993). Five Papers on WordNet, *Technical report*, Cognitive Science Laboratory, Princeton University. Revised version.

Oxford Reference Online (1999). that: pron, adj, adv, conj", *The Oxford American Dictionary of Current English*, Oxford Reference Online, Oxford University Press. 5 November 2004.

Resnik, P. and Smith, N. A. (2003). The Web as a Parallel Corpus, *Computational Linguistics* **29**(3): 349–380.

Sadat, F., Déjean, H. and Gaussier, . (2002). A Combination of Models for Bilingual Lexicon Extraction from Comparable Corpora, *In Proceedings of Papillon 2002 Seminar*, Tokyo, Japan, pp. 16–18.

Savický, P. and Hlavácová, J. (2002). Measures of Word Commonness, *Journal of Quantitative Linguistics* **9**(3): 215–231.

Sågvall Hein, A. (1998). De morfologiska beskrivningarna i Sve.Ucp., *Technical report*, Institutionen för lingvistik, Uppsala Universitet, Uppsala.

Sågvall Hein, A. (2004). Machine translation, Slides for "Maskinöversättning och språkgransking".

Somers, H. (1999). Knowledge Extraction from Bilingual Corpora, *in* M. T. Pazienza (ed.), *Information Extraction: Towards Scalable, Adaptable Systems*, Vol. 1714 of *Lecture Notes in Artificial Intelligence*, Springer, pp. 95–119. 2nd School on Information Extraction, SCIE-99. Frascati, Italy, June 28 - July 2, 1999.

Stora engelsk-svenska ordboken (1989). *Stora engelsk-svenska ordboken*, Norstedts Förlag.

Tapanainen, P. and Järvinen, T. (1997). A non-projective dependency parser, *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP'97) ACL*, Washington, D.C., pp. 64–71.

Thorell, O. (1977). *Svensk grammatik*, Scandinavian University Books. Second edition.

Turcato, D., Popowich, F., Laurens, O. and McFetridge, P. (1998). Reuse of linguistic resources in MT, *First International Conference on Language Resources and Evaluation (LREC)*, Granada, Spain.

Yang, D.-H., Lee, I.-H. and Cantos, P. (2002). On the Corpus Size Needed for Compiling a Comprehensive Computational Lexicon by Automatic Lexical Acquisition, *Computers and the Humanities* **36**: 171–190.

Zhang, H., Huang, C. and Yu, S. (2004). Distributional Consistency : A General Method for Defining ACore Lexicon, *4th International Conference on Language Resources and Evaluation(LREC2004)*, Lisbon. Portugal.

Appendix

A Morphological categories

A.1 The English entries

```
***Nouns***

cat-name stem s-nom pl-nom s-gen pl-gen

-s -+ +s +’s +s’
boy boy boys boy’s boys’

-es -+ +es +’s +es’
box box boxes box’s boxes’

-ies -y +y +ies +y’s +ies’
cit city cities city’s cities’

clothes -clothes -clothes’
man 0 man men man’s men’s
woman 0 woman women woman’s women’s
child 0 child children child’s children’s
basis 0 basis bases basis’s bases’
knife 0 knife knives knive’s knives’
life 0 life lives life’s lives’
self 0 self selves self’s selves’
wife 0 wife wives wife’s wives’
means 0 means means means’means’
foot 0 foot feet foot’s feet’s
money 0 money money money’s money’s


***Verbs***

**Auxiliary**

*Be*

base 1-sing 2-sing 3-sing ing 1-past past past-part

be am are is being was were been
```

```
*Other aux*

base 3-sing ing past past-part

have has having had had
do does doing did done

can can -could -
may may -might -
must must ---
ought to ought to ---
shall shall -should -
will will -would -

*Regular*

cat-name stem non-3-sing* 3-sing ing past

walk -walk walk+s walk+ing walk+ed
beg -beg beg+s beg+ging beg+ged (double consonant)
like -like like+s lik+ing lik+ed
pass -pass pass+es pass+ing pass+ed
carry -y carr+y carr+ies carry+ing carr+ied
lie -ie l+ie l+ies l+ying l+ied

*this form includes present tense with all persons (except third-person singular) a

*Irregular*

base past past-part pres-cat*

arise arose arisen like
awake awoke awaken like
bear bore borne walk
beat beat beaten walk
begin began begun beg
behold beheld beheld walk
bend bent bent walk
bet bet bet beg
bid bid bid beg
bind bound bound walk
bite bit bitten like
bleed bled bled walk
blow blew blown walk
break broke broken walk
breed bred bred walk
bring brought brought walk
build built built walk
burn burnt burnt walk
...
```

(only a sample of the irregular verbs are shown here)

*Forms in the present tense (i.e. 3-sing, non-3-sing, and -ing) are same as this re

Adjectives

cat-name base comp super

hard hard hard+er hard+est
big big big+ger big+gest
wide wide wide+r wide+st
dry dr+y dr+ier dr+iest
careful careful more careful most careful
good good better best
bad bad worse worst
far far further furthest
little little smaller smallest

Adverbs

cat stem pos comp sup

well -well bett+er best
badly -badly worse worst
much -much more most
hard -hard hard+er hard+est
close -close close+r close+st
early earl earl+y earl+ier earl+iest
far -far further furthest
far -far farther farthest

A.2 The Swedish entries

```
***Nouns***

cat-name stem pl-stem s-nom s-nom-def s-gen s-gen-def pl-nom pl-nom-def
pl-gen pl-gen-def gender

ros -0 + +en + +ens +or +orna +ors +ornas u
flicka -a 0 +a +an +as +ans +or +orna +ors +ornas u
stol -0 + +en +s +ens +ar +arna +ars +arnas u
hals -0 + +en + +ens +ar +arna +ars +arnas u
öken -en 0 +en +nen +ens +nens +nar +narna +nars +narnas u
afton -on 0 +on +onen +ons +onens +nar +narna +nars +narnas u
djävul -ul 0 +ul +ulen +uls +ulens +lar +larna +lars +larnas u
kam -0 + +men +s +mens +mar +marna +mars +marnas u
mun -0 + +nen +s +nens +nar +narna +nars +narnas u
fru -0 + +n +s +ns +ar +arna +ars +arnas u
gosse -e 0 +e +en +es +ens +ar +arna +ars +arnas u
nyckel -el 0 +el +eln +els +elns +lar +larna +lars +larnas u
seger -er 0 +er +ern +ers +erns +rar +rarna +rars +rarnas u
lämmel -mel 0 +mel +meln +mels +elns +lar +larna +lars +larnas u
hummer -mer 0 +mer +mern +mers +merns +rar +rarna +rars +rarnas u
finger -er 0 +er +ret +ers +rets +rar +rarna +rars +rarnas n
botten -en 0 +en +nen +ens +nens +nar +narna +nars +narnas u
sommar -mar 0 +mar +maren +mars +marens +rar +rarna +rars +rarnas u
moder -mödr + +n +s +ns +rar +rarna +rars +rarnas u
film film + 0 +en +s +ens +er +erna +ers +ernas u
plats plats + 0 +en + +ens +er +erna +ers +ernas u
vän vän + 0 +nen +s +nens +ner +nerna +ners +nernas u
nöt nöt + 0 +en +s +ens +ter +terna +ters +ternas u
vin vin + 0 +et +s +ets +er +erna +ers +ernas n
nivå nivå + 0 +n +s +ns +er +erna +ers +ernas u
historia -a 0 +a +an +as +ans +er +erna +ers +ernas u
möbel möb -el 0 +el +eln +els +elns +ler +lerna +lers +lernas u
drama dram -a 0 +a +at +as +ats +er +erna +ers +ernas n
strand strand -sträng + +en +s +ens +er +erna +ers +ernas u
brand -bränd + +en +s +ens +er +erna +ers +ernas u
stad -städ + +en +s +ens +er +erna +ers +ernas u
land -länd + +et +s +ets +er +erna +ers +ernas n
kläder 0 kläd 0 0 0 0 +er +erna +ers +ernas pl
pengar 0 peng 0 0 0 0 +ar +arna +ars +arnas pl
sko -0 + +n +s +ns +r +rna +rs +rnas u
fängelse -0 + +t +s +ts +r +rna +rs +ernas n
bonde -bönd 0 + +n +s +ns +er +erna +ers +ernas u
fot -fött 0 + +en +s +ens +er +erna +ers +ernas u
samhälle -0 +e +et +es +ets +en +ena +ens +enas n
bi -0 + +et +s +ets +n +na +ns +nas n
knä -0 + +t +s +ts +n +na +ns +nas n
bord -0 + +et +s +ets + +en +s +ens n
```

```
hus -0 + +et + +ets + +en + +ens n
huvud -et + +et +s +ets +en +ena +ens +enas n
gisslan -0 + + +s +s + + +s +s u
fader -0 fäd + +n +s +ns +er +erna +ers +ernas u
ordförande -0 + +n +s +ns + +na +s +nas u
garage -0 + +t +s +ts + +n +s +ns n
mil -0 + +en +s +ens + +en +s +ens u
kypare -e 0 +e +en +es +ens +e +na +s +nas u
tecken -en 0 +en +net +ens +nets +en +nen +ens +nenas n
program -0 + +met +s +mets + +men +s +mens n
tum -0 + +men +s +mens + +men +s +mens u
segel -el 0 +el +let +els +lets +el +len +els +lens n
fönster -er 0 +er +ret +ers +rets +er +ren +ers +rens n
nummer -mer 0 +mer +ret +mers +rets +mer +ren +mers +rens n
ögon -a 0 +a +at +as +ats +on +onen +ons +onens n
gås -gäss + +en +s +ens + +en + +ens u
man -män + +nen +s +nens + +nen +s +nens u
narkotikum -um 0 +um +umet +ums +umets +a +an +as +ans n
decennium -um 0 +um +umet +ums +umets +er +erna +ers +ernas n
prestanda -a 0 +a +an +as +ans +a +an +as +ans u
broiler -0 + +n +s +ns +s 0 +s 0 u
grädde -0 + +n +s +ns 0 0 0 0 u
mjölk -0 + +en +s +ens 0 0 0 0 u
kammare -mare 0 +mare +maren +mares +marens +mare +rarna +mares +rarnas
u
ledamot -0 ledamöt + +en +s +ens +er +erna +ers +ernas u
väsen -0 + +et +s +ets + +a +s +as n
förflutet -et 0 +et +na +ets +nas +na +nas +na +nas n
vuxen -en 0 +en +na +ens +nas +na +nas +na +nas u


Referens: Sågvall Hein, A. 1998. De morfologiska beskrivningarna i Sve.Ucp. Institu


****Verbs****

cat-name stem pres pret sup perf-part

*1st conjugation*

ropa ropa ropa+r ropa+de ropa+t ropa+d

*2nd conjugation*

stänga stäng stäng+er stäng+de stäng+t stäng+d
köpa köp köp+er köp+te köp+t köp+t
känna kän kän+ner kän+de kän+t kän+d
glömma glöm glöm+mer glöm+de glöm+t glöm+d
reda re re+der re+dde re+tt re+dd
mäta mä mä+ter mä+tte mä+tt mä+tt
```

vända vän vän+der vän+de vän+t vän+d
lyfta lyf lyf+ter lyf+te lyf+t lyf+t
köra kör kör kör+de kör+t kör+d

3rd conjugation

sy sy sy+r sy+dde sy+tt sy+dd
sy flå flå+r flå+dde flå+tt flå+dd

4th conjugation (irregular, including aux)

base pres pret sup perf-part

anlända anländer anlände anlänt anländ
använda använder använde använt använd
be/bedja ber bad bett bedd
binda binder band bundit bunden
bita biter bet bitit biten
bjuda bjuder bjöd bjudit bjuden
bli/bliva blir/bliver blev blivit bliven
bringa bringar bragte bragt bragd
brinna brinner brann brunnit brunnen
brista brister brast brustit brusten
bryta bryter bröt brutit bruten
byta byter bytte bytt bytt
bära bär bar burit buren
böra bör borde bort 0
dimpa dimper damp dumpit 0
dra drar drog dragit dragen
dricka dricker drack druckit drucken
driva driver drev drivit driven
drypa dryper dröp drupit/drypt 0
duga duger dög/dugde dugt 0
dyka dyker dök/dykte dykt dykt
dö dör dog dött 0
...
(only a sample of the irregular categories are shown here)

deponens

andas anda anda+s anda+des anda+ts -

particle verbs

cat+part base pres pret sup perf

tala+part tala om talar om talade om talat om talad om

reflexive verbs

uttala sig

Adjectives

cat stem utr neutr plur* best-mask komp super sup-best

kall -kall kall+t kall+a kall+e kall+are kall+ast kall+aste
hård -hård hår+t hård+a hård+e hård+are hård+ast hård+aste
ny -ny ny+tt ny+a ny+e ny+are ny+ast ny+aste
god -d go+d go+tt go+da go+de go+dare go+dast go+daste
våt -våt+t våt+a våt+e våt+are våt+ast våt+aste
kort -kort kort kort+a kort+e kort+are kort+ast kort+aste
dum -dum dum+t dum+ma dum+me dum+mare dum+mast dum+maste
sann -n san+n san+t san+na san+ne san+nare san+nast san+naste

dum(m) -dum dum+t dum+ma dum+me mer dum mest dum mest dum+ma
allmän -allmän allmän+t allmän+na allmän+ne mer allmän mest allmän mest allmän+na
kall(m) -kall kall+t kall+a kall+e mer kall mest kall mest kall+a
användbar -användbar användbar+t användbar+a användbar+e mer användbar
mest andvändbar mest användbar+a
kort(m) -kort kort kort+a kort+e mer kort mest kort mest kort+a
direkt -direkt direkt direkt+a direkt+e mer direkt mest direkt mest direkt+a
hård(m) -d hår+d hår+t hår+da hår+de mer hår+d mest hår+d mest hår+da
ny(m) -ny ny+tt ny+a ny+e mer ny mest ny mest ny+a
god(m) -d go+d go+tt go+da go+de mer go+d mest go+d mest go+da
våt(m) -våt våt+t våt+a våt+e mer våt mest våt mest våt+a

älskad -d älska+d älska+t älska+de älska+de mer älska+d mest älska+d mest älska+de
avsides -avsides avsides avsides avsides mer avsides mest avsides mest avsides
kul -kul 0 0 0 mer kul mest kul 0
hög -hög hög+t hög+a hög+e hög+re hög+st hög+ste
enda -a end+a end+a end+a end+e 0 0 0
egen -en eg+en eg+et eg+na eg+ne mer eg+en mest eg+en mest eg+na
enkel -el enk+el enk+elt enk+la enk+le enk+lare enk+last enk+laste
säker -er säk+er säk+ert säk+ra säk+re säk+rare säk+rast säk+raste
trogen -en trog+et trog+na trog+ne trog+nare trog+nast trog+naste

stor stor stor+t stora store större störst störst+e/+a
låg låg lågt låga låge lägre lägst lägst+e/+a
lång lång långt långa långe längre längst längst+e/+a
ung ung ungt unga unge yngre yngst yngst+a/+e
grov grov grovt grova grove grövre grövst grövst+a/+e
nära nära nära nära nära närmare närmast närm+aste/+msta
in/inne inne inne inne inne inre innerst innersta
bort (AD) bortre borterst bortersta
under (PREP) undre underst understa
bra bra bra bra bra bättre bäst bäst+a/+e
dålig dålig dålig+t dåliga dålige sämre sämst sämst+a/+e
gammal gammal gammalt gamla gamle äldre äldst äldst+a/+e

många 0 0 många 0 fler flest flesta
få 0 0 få 0 färre 0 0
mycken mycken mycket myckna myckne mer/+a mest mesta
liten liten litet små lill+e mindre minst minst+e/+a
sist 0 0 0 0 0 sist siste
rädd 0 rädda rädde räddare räddast räddaste
viss viss visst vissa visse 0 0 0 0

*This form is both indef plural and all def forms, except maskulinum

Adverb

cat stem pos comp sup

snabbt snabb+t snabb+are snabb+are
fort fort+ fort+are fort+ast
sent sen+t sen+are sen+ast
länge läng+e läng+re läng+st
ofta ofta+ ofta+re ofta+st
före för+e för+r för+st

mycket mycket mer mest
gärna gärna hellre helst
illa illa värre värst
illa illa sämre sämst
dåligt dåligt sämre sämst
bra bra bättre bäst
nära nära närmare närmast

B Corpus Statistics

Statistics of the project corpus

abbreviations

e-words=english words (exkl num, exkl interp)
s-words=Swedish words
+punct=tokens including punctuations
sent=sentences
to/s=tokens per sentence
t-pairs=types of linked lemma pairs
to/ty=tokens per type

sample	a95	aXP	be	g	hp	ep98	ep00	sA1	sA2	sB
sent	400	400	400	400	400	230	320	500	500	500
e-token	5342	6159	5680	7901	5203	5640	8204	3262	2838	2359
e-types	657	768	1654	1918	1147	1045	1338	788	572	461
e-to/ty	8,13	8,02	3,43	4,12	4,54	5,40	6,13	4,14	4,96	5,12
e-to/s	13,4	15,4	14,2	19,75	13,01	24,52	25,64	6,54	5,68	4,72
s-token	4698	5299	5483	8110	5364	5166	7254	2812	2197	1849
s-types	741	923	1850	2257	1417	1182	1524	899	702	498
s-to/ty	6,36	5,74	2,96	3,59	3,79	4,37	4,76	3,13	3,13	3,71
s-to/s	11,7	13,25	13,71	20,28	13,41	22,46	22,67	5,62	4,39	3,70
l-pairs	1313	1751	2213	2905	1927	1718	2504	1198	980	696

Number of nouns

Eng: 16124
Swe: 14600

Number of verbs

Eng:
V = 8320
ING = 1070
EN = 1561
tot = 10951

Swe:
V = 8657
AD = 667

tot = 9324

Number of adjectives

Eng: 3576
Swe: 3630

Number of adverbs

End: 3154
Swe: 3473

Nr of NDE (Swedish)

411

Number of English function words

PREP 6265
CC 1929
CS 978
NEG-PART 471
INFMARK 942
DET 6573
tot = 17158

C Example from WordNet

'Extremity', with all its hyponyms:

extremity, appendage, member – (an external body part that projects from the body; "it is important to keep the extremities warm")

→ chelicera – (either of the first pair of fang-like appendages near the mouth of an arachnid; often modified for grasping and piercing)

→ mouthpart – (any part of the mouth of an insect or other arthropod especially one adapted to a specific way of feeding)

→ claw, chela, nipper, pincer – (a structure like a pincer on the limb of a crustacean or other arthropods)

→ parapodium – (one of a pair of fleshy appendages of a polychete annelid that functions in locomotion and breathing)

→ fin – (organ of locomotion and balance in fishes and some other aquatic animals)
→→ dorsal fin – (unpaired median fin on the backs of fishes and some other aquatic vertebrates that help to maintain balance)
→→ pectoral fin – (either of a pair of fins situated just behind the head in fishes that help control the direction of movement)
→→ pelvic fin, ventral fin – (either of a pair of fins attached to the pelvic girdle in fishes that help control the direction of movement; correspond to hind limbs of a land vertebrate)
→→ tail fin, caudal fin – (the tail of fishes and some other aquatic vertebrates)
→→→ heterocercal fin – (tail fin with unequal lobes in which the vertebral column turns upward into the larger lobe as in sharks)
→→→ homocercal fin – (symmetrical tail fin extending beyond the end of the vertebral column as in most bony fishes)

→ swimmeret, pleopod – (one of the paired abdominal appendages of certain aquatic crustaceans that function primarily for carrying the eggs in females and are usually adapted for swimming)

→ limb – (one of the jointed appendages of an animal used for locomotion or grasping: arm; leg; wing; flipper)
→→ hind limb – (a posterior leg or homologous structure in other animals)
→→→ hind leg – (the back limb of a quadruped)
→→ forelimb – (the front limb (or homologous structure in other animals such as a flipper or wing))
→→→ foreleg – (the front limb of a quadruped)

→→ flipper – (the flat broad limb of aquatic animals specialized for swimming)
→→ leg – (a human limb; commonly used to refer to a whole limb but technically only the part between the knee and ankle)
→→→ pin, peg, stick – (informal terms of the leg; ”fever left him weak on his sticks”)
→→→ bowleg, genu varum, tibia vara – (a leg bowed outward at the knee (or below the knee))
→→→ shank’s mare, shanks’ mare, shank’s pony, shanks’ pony – (you own legs; ”I traveled on shank’s mare”)
→→→ spindlelegs, spindleshanks – (long thin legs)
→→→ knock-knee, genu valgum, tibia valga – (inward slant of the thigh)
→→ crus – (the leg from the knee to foot)
→→ leg – (a structure in animals that is similar to a human leg and used for locomotion)
→→→ animal leg – (the leg of an animal)
→→ thigh – (the part of the leg between the hip and the knee)
→→→ lap – (the upper side of the thighs of a seated person; ”he picked up the little girl and plopped her down in his lap”)
→→ arm – (a human limb; technically the part of the superior limb between the shoulder and the elbow but commonly used to refer to the whole superior limb)
→→ cubitus – (the arm from the elbow to the fingertips)
→→ forearm – (the part of the superior limb between the elbow and the wrist)

→ **vertebrate foot**, pedal extremity – (the extremity of the limb in vertebrates)
→→ animal foot, foot – (a foot of a vertebrate other than a human being)
→→→ fossorial foot – (foot adapted for digging as in moles)
→→→ hoof – (the foot of an ungulate mammal)
→→→→ cloven foot, cloven hoof – (a hoof divided into two parts at its distal extremity (as of ruminants or swine))
→→→→ horse’s foot, horse’s hoof – (the hoof of a horse)
→→→ bird’s foot – (the foot of a bird)
→→→→ claw – (a bird’s foot that has claws)
→→→→ zygodactyl foot – (having the first and fourth toes of each foot directed backward and the second and third forward)
→→→→ heterodactyl foot – (having the first and second toes of each foot directed backward and the third and fourth forward)
→→→→ webbed foot – (a bird’s foot with folds of skin between the toes)
→→→→ lobate foot – (a foot having separate toes each with membranous flaps along the sides)
→→→ webfoot – (a foot having the toes connected by folds of skin)
→→→ trotter – (foot of a pig or sheep especially one used as food)
→→→ forefoot – (a front foot of a quadruped)
→→→ hindfoot – (a rear foot of a quadruped)
→→→ paw – (a clawed foot of an animal especially a quadruped)
→→→→ forepaw – (front paw; analogous to the human hand)
→→→→→ hand – (terminal part of the forelimb in certain vertebrates (e.g. apes or kangaroos)
; ”the kangaroo’s forearms seem undeveloped but the powerful five-fingered hands are skilled at feinting and clouting”- Springfield (Mass.) Union)
→→ foot, human foot, pes – (the foot of a human being; ”his bare feet projected from his trousers”; ”armored from head to foot”)
→→→ flatfoot, splayfoot, pes planus – (a foot afflicted with a fallen arch; abnormally flattened and spread out)

→ digit, dactyl – (a finger or toe in human beings or corresponding part in other vertebrates)

→→ finger – (any of the terminal members of the hand (sometimes excepting the thumb); ”her fingers were long and thin”)
→→→ thumb, pollex – (the thick short innermost digit of the forelimb)
→→→ index, index finger, forefinger – (the finger next to the thumb)
→→→ ring finger, annualry – (the third finger (especially of the left hand))
→→→ middle finger – (the second finger; between the index finger and the ring finger)
→→→ pinkie, pinky, little finger – (the finger farthest from the thumb)
→→ toe – (one of the digits of the foot)
→→→ big toe, great toe, hallux – (the first largest innermost toe)
→→→ hammertoe – (a deformed toe which is bent in a clawlike arch)
→→→ little toe – (the fifth smallest outermost toe)