Domain specific corpus building and lemma selection in a computational lexicon

Stig W. Jørgensen, Carsten Hansen, Jette Drost
Dept. of Computational Linguistics, Copenhagen Business School,
Bernhard Bangs Allé 17B, DK-2000, Denmark.
{swj.id, ch.id, jd.fra}@cbs.dk

Dorte Haltrup, Anna Braasch, Sussi Olsen
Centre for Language Technology, Njalsgade 80, DK-2300, Denmark.
{dorte, anna, sussi}@cst.dk

1. Introduction

The Danish STO project, SprogTeknologisk Ordbase (i.e. Lexical database for language technology) is a national project with the aim of creating a large size Danish lexicon for natural language processing. This work greatly benefits from the Danish lexicon material developed within the framework of the multinational PAROLE project (see LE-PAROLE 1998).

Most of the researchers working on the project are computational linguists, but the group also includes general linguists, lexicographers and people with a background in language for specific purposes (LSP). The work reported in this paper is very much a result of this combination of various expertises.

The STO lexicon is planned to contain a minimum of 50,000 lemmas, of which approximately 15,000 will come from six different domains of language for specific purposes. The lemmas are provided with detailed morphological and syntactic information while the semantic information in this first version for a large part of the vocabulary is reduced to domain information. The planned number has already been reached for the encoding of morphological descriptions; 24,000 entries have also been encoded with syntactic descriptions.

The first version of the lexicon should be complete by the end of 2003, but extension of the linguistic information especially at the semantic level, addition of pronunciation information as well as an extension of the number of domains will be possible at a later date, depending on funding. For further information on the STO project, see Braasch (Braasch et al. 1998, Braasch 2002) and the STO homepage: www.cst.dk/sto/uk.

2. The domains and corresponding corpora

The selection of specific domains is based on the potential future applications. At present, work has been done on the domains of IT (information technology), environment and health, and is just starting on the domain of public administration. The subjects of the two remaining domains have not been decided.

Existing lemma lists or dictionaries from the different domains are not directly suitable for our purpose, for various reasons.

Firstly, the delimitation of a domain varies a lot, e.g. the domain IT in some lemma lists includes lemmas from domains like commerce and marketing, and some health domain lemma lists have a large proportion of lemmas relating to nutrition, bodily exercise and alternative medicine, while other lemma lists have a more narrow definition merely including technical or scientific terms.

Secondly, lemma lists and domain specific dictionaries are made for different purposes and address different user profiles. Mostly they are highly specialised term lists, while some cover a broader vocabulary like the one appearing in newspapers or words from general language having another meaning in the language of the domain in question. The STO database is not supposed to cover the most specialised terms of the different domains but rather the vocabulary that laymen might encounter in various contexts. Specialised termlists can later be added by future users.

Thirdly, Danish is a less widely spoken language and up-to-date lemma lists or dictionaries are not available for all domains.

For these reasons, the lemma selection for the domain specific vocabulary of STO is primarily based on text corpora. As a matter of fact this goes for the STO project as such. STO is supposed to be corpus based, which means that not only the lemma selection for both general language and domain specific language but also the morphological, syntactic and semantic encoding of each lemma depends on what is found in the relevant corpora.

The selection of specific domains and the lemma selection from these was one of the first tasks to be initiated in the project (previously reported in Olsen 2002), but the procedure for establishing corpora and selecting lemmas is being continuously refined. We are aware that the vocabulary of the domains changes very fast and that the vocabulary of the STO database will be out of date after a short time. But since we have established a method, and the process of collecting texts for the corpora and the editing of the lemma candidate lists is to a large degree automatic, it will be rather unproblematic to extend the vocabulary of each domain in the future and thus keep it up-to-date.

At this moment we have built three corpora, IT, environment and health. The health corpus was built from scratch according to the methods described below, whereas the other two corpora were initiated at an earlier stage of the project and have subsequently been revised to conform to the standards of the renewed methodology. From the IT domain we have selected and encoded – morphologically and syntactically – about 2000 words, and a similar number from the environment domain, whereas the health domain turned out to be richer with approximately 2500 lemmas.

3. The building of domain specific corpora

The main source for texts in the corpora is the web, since WWW texts are easily available and can be collected with a high degree of automatisation. Using the web as a resource for lexicography poses certain problems, including the fact that is is often a “dirty” source with no guaranteed level of correctness with respect to spelling and grammaticality (see Grefenstette 2002 for a discussion of webbased lexicography). The most important problem, however, is the selection of the texts to be included in a given corpus.

There is a danger of circularity in the process of establishing a domain corpus. The purpose of the corpus is to discover which words are typical and frequent in a given domain, but to establish a representative corpus we need a notion of the delimitation and composition of the domain. If the corpus is to be built from available web pages, relevant search words have to be used to identify useable websites. In a sense one needs to know some of the terms in advance, even if the end product of the task is to discover these lemmas! It is not possible to avoid this basic problem altogether, but it is possible to develop a strategy which neither results in a purely arbitrary composition of the corpus, nor falls into the opposite trap by giving it such a fine-grained structure that the result is more or less determined in advance.

Our strategy is an onomasiological approach[1] where what we call a *onomasiological structure*, or OS, serves as definition and delimitation of the domain, and our main goal is a corpus with sufficient *coverage* of the domain.

The corpus building process consists of four main tasks, as sketched in table 1:

Task 1	Construction of OS for the domain
Task 2	Identification of relevant websites
Task 3	Downloading and sorting of text
Task 4	Editing text into a corpus

Table 1

Though there is a natural progression from task 1 to task 4, these should not simply be read as steps in the process: Once a first version of the OS is established, tasks 1-3 are performed more or less simultaneously, with mutual feedback.

Task 1: On the basis of existing thesauri and available literature, including major encyclopedias, we construct an onomasiological structure – the OS – a hierarchically structured list of topics and key words relating to a domain. That means that the OS is constructed very much like a thesaurus, but it does not (necessarily) fulfil the formal requirements of a thesaurus (or a concept system, for that matter, in spite of the fact that the OS is often presented graphically like a concept system). A thesaurus is “a structured list of expressions intended to represent in unambiguous fashion the conceptual content of the documents in a documentary system and of the queries addressed to that system” (Eurovoc homepage 2002). The OS serves a different purpose: to establish the collection of documents.

Task 2: The items from the OS are used as search terms on the web to identify relevant texts. The search is performed using standard search engines, and with the appropriate amount of creativity usually employed by experienced web users (substituting search words for related words, trying various spelling variants etc.). Relevant texts are those that cover one, or preferably more, topics of the domain and confirm to the general standards of the project, most importantly:

- Language: Danish. (A trivial demand, but often a practical problem, since many Danish websites also include text in English)
- Time: Modern. We want the corpora to be as up-to-date as possible.
- Communication: Preferably expert to layman, not expert to expert. The lemmas we wish to extract “should not be highly specialized terms of the domain, but rather words that laymen

[1] The onomasiological approach is a principle of the so-called Vienna school of terminology. Terminology studies concepts before terms. In this tradition, unlike in lexicology, onomasiology “does not refer to the content aspect of the sign but rather to the concept seen as part of the world outside language” (Temmerman 2000: 5).

> have to read and understand as part of their everyday life" (Braasch 2002). This also corresponds to the fact that we use a newspaper corpus as a main resource for the general language part of the lexicon.

If the seach reveals new relevant topics not covered by the OS, or if the search yields insufficient or inappropriate results, the searcher reports to the OS constructor who improves or extends the structure. Relevant web addresses are passed on to task 3.

Task 3: Texts are downloaded from the web pages. HTML codes are deleted, and foreign (primarily English) documents and identical documents are removed. These processes are performed automatically and can pose certain technical problems. New and constantly changing web technology (advanced Java scripts etc.) may render the extraction of text difficult with the available tools, and the removal of documents containing foreign text is often less than perfect when done automatically, but very time-consuming if done manually. If the net result (in terms of size of the resulting text file) is poor, we cannot guarantee that the body of text covers the topics it is supposed to cover. We do not engage in the time-consuming labour of reading through the text, but report back to task 2 that the web page in question does not ensure coverage, that is, we need more relevant web addresses on a given topic.

Task 4: Once coverage is achieved, the texts are assembled into a corpus with a minimum of editing. In rare cases, a downloaded text is reduced in size before entering into the corpus. This is only done in extreme cases where a website has been included to ensure coverage of a minor and relatively unimportant topic, but the resulting text file is inappropriately huge compared to the rest of the texts and would distort the statistics of the lemma selection list that is to be derived from the corpus. Thus, reduction is only used with great caution, in accordance with our pragmatic goal of ensuring coverage, but not weighting.

An important feature of this approach is that we have not predefined a size for the corpus we are building. Rather, the corpus is of sufficient size when it is of sufficient coverage! This leaves room for individual differences depending on the specific domain. It is our experience that the approach results in corpora in the range between 1 and 2 million tokens, which has proved appropriate to our purposes and the lemma selection methods.

A major problem with the approach is that the actual composition of a given corpus depends very much on our tools. This has to do with the technical difficulties mentioned in task 3. If we could simply download all the text we wanted, the process would be much simpler with less feedback between the different tasks. As it is now, we are locked in an arms race with the web designers, and either we have to spend more time to acquire or produce new tools that can extract text from more advanced web pages, or we have to spend the time, as we do now, to find websites that our tools can handle.

4. The selection of domain specific lemmas

The overall method for the lemma selection is sketched in table 2:

Step 1	Tokenisation (and POS-tagging of corpus)
Step 2	Lemmatisation
Step 3	Generation of lemma candidate list
Step 4	Manual examination of lemma candidates
Step 5	Quality evaluation

Table 2

The method has been changed recently, mainly by the introduction of a new lemmatiser. What we describe here is the new method (which was applied in the health domain), with a few remarks on the methods that were used previously.

Step 1: Firstly, the domain specific corpus is tokenised, involving separation of punctuation marks from word forms and identification of abbreviations and common multiword units. The tokens are part-of-speech tagged, using the Brill tagger.

Step 2: Previously, normalisation of the word forms was performed in a separate step, that is, capitalisation and other special symbols like the special Danish characters were substituted by lower-case letters or other characters respectively. As the next step, the normalised words forms were truncated according to manually constructed rules (based on a list of the 35 most common Danish word endings and their contexts), and the truncated forms were grouped together as candidate lemmas. This truncation and grouping of word forms was used as a temporary step since we did not have access to a lemmatiser for Danish, capable of treating unknown words. Now the truncation has been discarded, and normalisation is performed in combination with lemmatisation by a lemmatiser that is currently being developed at the Centre for Language Technology.

The lemmatiser consists of the following linguistic components: i) a lexicon containing 450.000 word forms with their corresponding 60.000 lemmas and morpho-syntactic categories, ii) 43.000 rules generated automatically from a list of word forms and corresponding lemmas, and iii) frequencies of word forms and lemmas from other corpora, used in the disambiguation process of homographs. In the present case, the input to the lemmatiser is the list of part-of-speech tagged word forms from the collected corpus. Each token is looked up in the lexicon, according to the following principles:

```
IF it is unambiguous then the lemma is selected,
IF it is ambiguous then the input part-of-speech tag is used to disambiguate, but
    IF the homographs are within the same word class
    then frequencies of lemmas and word forms are used to disambiguate, or
    IF the part-of-speech tag and the morpho-syntactic category in the lexicon do not agree,
    then the lexicon overrules
IF the token is unknown to the lexicon then rules are applied to generate the appropriate lemma
from the word form.
```

The lemmatiser has different possible output formats. For the lemma selection the output is a list of: lemmas, lemma frequencies, frequencies of the word forms in the input, part-of-speech tags, and marking of word forms not found in the lexicon and of word forms found with another part-of-speech tag in the lexicon.

The lemmatiser has a performance rate of 97.8% correct lemmas for lemmatisation with dictionary and part-of-speech tagged input. However, this result depends on the part-of-speech tagging being correct; the actual outcome in the project is less perfect, since the Brill tagger introduces some errors, mostly relating to proper names. The performance rate for lemmatisation with untagged input is 94.5% correct lemmas. We therefore consider doing lemmatisation without the part-of-speech tagging.

Step 3: The dictionary used during the lemmatisation consists of the words already encoded in the STO database. New lemmas not found in the dictionary are labelled as such, which enables us to put them on a separate list. The list is then sorted by frequency. This list of lemmas not previously encoded is what we call the lemma candidate list.

Step 4: The lemma candidate list is examined manually, ignoring lemmas with a frequency of 1 (i.e., only a single occurrence in the domain corpus). The main purpose of the manual examination is to select the actual domain lemmas from the candidates (see section 5 below). Of course, the examination also includes correcting errors of lemmatisation and part-of-speech tagging. Previously, the part-of-speech assignment was made manually during this step, and we may return to this if we decide to do lemmatisation without part-of-speech tagging, but we also experiment with having only nouns, verbs and adjectives on the candidate list.

A couple of figures to illustrate the outcome of the process: The health corpus consisted of approximately 2 million tokens. This resulted in a candidate list containing approximately 17,000 lemmas, of which 7000 had a frequency higher than 1 (3000 had a frequency higher than 3). From the 7000 candidates with more than one occurrence, 2200 domain lemmas were selected. (An additional 600 were put on a list of general language words – see section 5 below).

Step 5: Finally, a quality evaluation is performed on the resulting vocabulary. One way of doing this is by comparing it to existing dictionaries or lemma lists of the domain. Though we do not find such dictionaries suitable as a primary source for our vocabulary, we can use them to check whether some central words of the domain accidentally do not figure in the selected vocabulary. The selection from the environment domain was compared to the *Miljøordbog*, an environmental dictionary (Heise and Kaalund 2001), and included every single item of this term list. In relation to the health domain, we are performing an experiment that will also serve as a quality evaluation of our corpus building method. We have acquired a body of health texts from the Danish National Encyclopedia. We will use these texts as an additional health corpus, treat it to the same methods, and see if the resulting lemma list is significantly different from the one we obtained from our webbased health corpus. In general, we are also experimenting with varying the corpus size (that is, including more than our method prescribes) and/or including more low-frequency lemmas to see how this changes the final result.

5. Principles of the manual selection

The manual selection of lemmas from the lemma candidate list is performed simply by deleting inadequate candidates. The following are deleted from the list:

- Proper names. All names of persons are deleted, with no respect to their frequency. Names of important organisations and institutions are preserved if sufficiently frequent. In the domain of environment, for instance, a name like Greenpeace is included, but the name of the Danish Environment Secretary is not.
- Expert terms. Terminology more appropriate of expert-to-expert communication than the layman-oriented vocabulary we wish to cover. For instance, formulas and long chemical names are deleted.
- Long and/or unusual compounds. In Danish orthography, like in German, but unlike English, compounds are written as single words, which accentuates the problem. It is viewed as more important to cover all relevant simplex words from which the compounds can be generated. Furthermore, unusual compounds are often instances of what we call expert terms above.
- Overrepresentation. Since our corpora are constructed on the basis of coverage, not weight, it is possible that lemmas from one or more less important subdomains become too dominant on the list. These are deleted if the overrepresentation is obvious from reading through the list. Of course, some basic steps have been taken to prevent distortion: We delete identical documents from the corpus, and reduce the size of documents when necessary (see section 3, tasks 3 and 4, above). Furthermore, we plan to differentiate the frequencies given in the lemma candidate list, so that the figures will not just reflect the frequency of a given lemma in the corpus as a whole, but also show how the occurences are distributed among the different text sources.
- Errors. Typos etc., or errors due to the automatic lemmatisation. Orthographical errors relating to compounding (Danish compounds are often, mistakenly, written as more than one word, which can lead to unlikely lemma candidates). Non-Danish words may occur due to foreign language quotes in the corpus texts. One should, however, hesitate to delete foreign words since a word from a foreign language may be used as a technical term in Danish. In some domains, a "foreign" spelling (e.g. "ch" for "k" in names of chemicals) is usual, even if it is unauthorized.

Of course, the list will not just include domain specific words, but also words belonging to general vocabulary that are not in the database already. Words that seem common are not deleted, but put on a separate list to be included among the general language lemmas.

6. Some specific problems

In this section we will discuss two very different, specific problems we have encountered in our work, to illustrate the kind of issues that can be raised by our methods and procedures. The first problem is one of automatic lemma selection. The procedure is not perfect, and there is a need for additional methods, for instance for the selection of collocations and multi-word units. Here, we will address another kind of problem, namely how to detect those words that appear both in the general language vocabulary and in a specific domain. The second problem is of a more theoretical nature and has to do with our corpus building methods. We identify WWW texts that deal with specific topics, but "text on the web" is not a homogenous genre, and the type of text very much depends on the domain. Does that create too large a typological difference between two given domain corpora?

6.1. Lemmas occurring in both general language and domain specific language

Words that appear in the general language vocabulary of STO will automatically be excluded from a candidate list of domain specific lemmas according to step 3 of the lemma selection described above. But a word might belong to the general language and at the same time be part of the language of a specific domain, so we need a method to detect words that have already been encoded as general language words but are also found with another meaning and perhaps another syntactic and/or morphological behaviour in a specific domain. Table 3 shows some examples.

Semantic difference		
Word	*Domain*	*Meaning*
'bus'	general language	a passenger vehicle
	IT	a data channel
'port'	general language	a large door or gate
	IT	an external computer connection
Morphological difference		
Word	*Domain*	*Plural inflexion*
'indeks' (Eng. 'index')	general language	'indekser'
	IT and mathematics	'indekser', 'indices'

Table 3

In order not to lose track of these lemmas, in the STO database we mark all entry words with source reference indicating in which corpus a lemma appears. Thus, the lemma 'bus' will be source-marked both for general language and IT since it appears in both corpora.

This means that a single word can have source reference to general language as well as to all specific domains. This will be the case for the most common general language words. Words from the general language with low frequency will (hopefully) only appear in the general language corpus and will not be object to further treatment, but words that are marked with source reference from a general language corpus and from one or two domain specific corpora have to be picked out for special treatment.

These lemmas will be subject to a special encoding process. For each lemma it has to be decided whether the linguistic behaviour of this lemma in each domain in which it occurs, differs from the existing encoding of the general language lemma at all the three linguistic levels. Any linguistic behaviour – morphological, syntactic or semantic – that differs from the general language encoding demands an encoding that reflects this behaviour and makes it apparent that the encoding is specific for a certain domain. Thus the 'indeks' example in table 3 will have two morphological units containing the inflectional patterns connected to it, one of which will be marked as valid for the IT domain only.

We have not yet started the process of encoding the lemmas with a particular, linguistic behaviour when appearing in a domain specific context. The process will be started when all domain corpora have been established.

6.2. Domain-dependant text type variations

In section 3 above we noted that our domain corpora should consist of texts that are in Danish, as up to date as possible, and preferably written by an expert for an audience of laymen. Without going into the details of text and corpus typology (Atkins et al. 1992), we can state that demands to the text define the corpus we are building, including of course the demand that the texts should fall within a given domain. But what about the decision to use the WWW as a source? A main reason for using the web is the pragmatic one already mentioned: WWW texts are easily available and can be collected with a high degree of automatisation – and they are up to date in the sense of being available now. One could however add a less pragmatic reason for using the web as a resource. Some of the potential applications that the STO lexicon could be used for, will deal with the web and web texts, so for this reason the web itself could be a suitable choice of resource. Our attitude towards the corpora we establish depends on whether we view the choice of resource as purely pragmatic or partly dictated by the nature of potential users and applications.

In comparing the environment corpus and the health domain corpus, we noted that most of the environment texts turned out to be texts originally produced in print (papers, reports, book chapters etc.) and subsequently made available on the web, whereas most of the health texts were produced for the web without prior publication (FAQs and online medical information services). Whether we should accept this result, depends on our attitude to WWW as a resource. If we maintain that we primarily gather our texts from the web to make life easier, we have failed in the sense that we have overlooked a parameter that is necessary to obtain uniform corpora – or, more precisely, that our criterion of "expert-to-layman communication" is too general to ensure uniformity. In that case, we should seek out other resources than the web, and add new kinds of material to the corpora when necessary. If, however, we want the corpora to be representative of the texts available on the WWW within given domains, the difference between the environment corpus and the health corpus does not have to be a problem. The composition of the corpora simply reflects the fact that web texts on environment typically take the form of papers and reports, whereas web texts on health do not, and web-based applications should benefit from this being reflected in our vocabulary.

Since the corpora used for the general language vocabulary are not web-based, there are good reasons for maintaining that the use of WWW is mainly pragmatic. Our present attitude is that we should simply pay more attention to this aspect of text collecting – and seek to improve our quality evaluation methods. As it is, we have no conclusive evidence of how the difference in text type actually affects the vocabulary. Our experiments with comparing lemma lists from web-based and non-web-based corpora within the same domain, may yield results that shed light on this question.

7. Concluding remarks

In this paper we have presented an approach to domain specific corpus building and lemma selection; an approach that tries to bring a certain amount of structure into the text selection process without restricting the outcome, and tries to apply a high degree of automatisation while still leaving some key decisions to human agents. The methodology is still being developed. We have been generally satisfied with our results, but we need to establish more rigorous procedures for quality evaluation, and we still await the outcome of the various experiments that have been suggested in this paper.

References

Atkins S, Clark J, Ostler N 1992 Corpus Design Criteria. *Literary and Linguistic Computing* 7(1): 1-16.

Braasch A 2002 Current developments of STO – the Danish Lexicon Project for NLP and HLT applications. In *Proceedings from the Third International Conference on Language Resources and Evaluation*, Las Palmas, pp 986-992.

Braasch A, Buhr-Christensen A, Olsen S, Pedersen B S 1998 A Large-Scale Lexicon for Danish in the Information Society. In *Proceedings from the First International Conference on Language Resources and Evaluation*, Granada, pp 249-255.

Eurovoc 2002 http://europa.eu.int/celex/eurovoc/ The European Communities.

Grefenstette G 2002 The WWW as a resource for lexicography. In Marie-Hélène Corréard (ed.) *Lexicography and Natural Language Processing. A Festschrift in Honour of B.T.S. Atkins.* Göteborg, EURALEX, pp 199-215.

Heise P, Kaalund L 2001 *Miljøordbog*. Copenhagen, Amtsrådsforeningen. Electronic version: www.miljoeordbog.dk

LE-PAROLE 1998 *Danish Lexicon Documentation.* Report. Copenhagen, Center for Sprogteknologi.

Olsen S 2002 Lemma selection in domain specific computational lexica – some specific problems. In *Proceedings from the Third International Conference on Language Resources and Evaluation*, Las Palmas, pp 1904-1908.

Temmerman R 2000 *Towards new ways of terminology description*. Amsterdam/Philadelphia, John Benjamins Publishing.