A Health Corpus Selected and Downloaded from the Web - Is it Healthy Enough?

Anna Braasch
Center for Sprogteknologi
University of Copenhagen
Njalsgade 80
DK-2300 Copenhagen S
DENMARK
anna@cst.dk

Abstract
The work on the Danish corpus-based computational lexicon STO was finished by the end of February 2004. This is the most comprehensive computational lexicon of Danish developed for NLP/HLT applications containing also lemmas from six different domains. The domain vocabulary has been selected from corpora created by assembling texts from the web; these corpora also form the basis for linguistic analysis/description. The average size of these corpora is 1.5 million tokens. Although the method of text selection and encoding of linguistic information was identical for all domains, a comparison of the lexicon entries showed that the health domain entries are considerably less complex than other domains. In order to investigate the cause of this difference, an additional corpus consisting of encyclopedic articles has been used for control purposes as regards the health domain. This paper focuses on a comparison of syntactic structures, esp. the variety of prepositional complements in these two corpora. Firstly, the basic properties of both corpora are discussed from the point of view of comparability; secondly, the objectives and the method of the comparison are outlined; thirdly some results and insights gathered from the work are presented. Finally, a method of using supplementary corpus look-ups in the case of deficiencies is suggested.

1 Introduction

The aim of the Danish STO project was to create a large lexicon for natural language processing. It contains over 81,000 lemmas, of which approx. 14,000 come from six different domains of language for specific purposes (LSP). The STO database is not intended to cover highly specialised terms but focuses on words of the domain languages "that laymen will have to read and understand as part of their everyday life." (Braasch, 2002). We consider this to be a kind of transitional area between the general language and specialised expert languages. The STO lexicon is corpus based both as regards the selection and the description of lemmas. All lemmas are provided with exhaustive descriptions of their inflectional properties, and 45,000 of them also with a fine-grained syntactic description as well. The linguistic descriptions are based on corpus analysis, and all lemmas are treated in a uniform way.

2 WWW Domain Corpora and Two Corpora of the Health Domain

2.1 Assembling domain texts from the WWW

For each of the six domains selected, a corpus was assembled with the web as text source. For resolving the basic problem: the selection of appropriate texts for the corpus, a kind of 'bootstrapping' strategy has been adopted. The method for establishing corpora and selecting lemmas is described in detail in Jørgensen et al. (2002) with the main point: "Our strategy is an onomasiological approach where what we call an *onomasiological structure*, or OS, serves as definition and delimitation of the domain (...) by means of the hierarchically ordered topic labels and key words of the domain." This method involves a risk of circularity if the labels of the taxonomy are both typical and frequent words of the domains and function as search words in identifying relevant homepages as well. In order to reduce this risk, each OS is constructed on the basis of classifications of the domain such as EUROVOC (Drost, 2002), also including the taxonomy worked out for the Danish National Encyclopedia, the so-called SDE-taxonomy.

2.2 Language investigations on www domain corpora

Although the method of text selection for the corpora, their analysis and the encoding of linguistic information were identical for all domains, a comparison of the lexicon entries showed that the health domain entries were considerably less complex than other domains. Table 1 gives a simplified overview of the complexity of syntactic frames of nouns from four different domains. There are two obvious criteria of comparison with regard to the syntactic complexity encoded in the lexicon. Firstly, the ratio of syntactic reading to lemma, viz. how many different complementation patterns are observed for one lemma; secondly, the distribution of syntactic descriptions onto valence frame types. The distribution column shows the number of nouns being zero- (0), mono- (1), bi- (2), tri- (3) or tetravalent (4), respectively.

Domain of the WWW corpus	No. of nouns	Syntactic readings	Noun/Reading Ratio	With Valence Total / %	Distribution on valence frames
Health *	1604	1615	1.006	98 / 6.1%	0= 1,517; 1= 94; 2= 4;
Administration	2430	2775	1.14	925 / 33.4%	0= 1,850; 1= 817; 2= 104; 3= 4; 4= 0
Finance	1259	1330	1.05	677 / 50.9%	0= 653; 1= 580; 2= 94; 3= 3; 4= 0
Commerce	1539	1802	1.17	795 / 44.2%	0= 1,007; 1= 670; 2=117; 3= 7; 4= 1

Table 1: Complexity of noun complementation recorded from the WWW corpora

This overview illustrates clearly the divergence of the syntactic complexity observed in the different www-corpora. In order to find the cause of this divergence, an additional corpus consisting of encyclopedia articles has been used for control purposes as regards the health domain. This raised two fundamental questions. Firstly, do the figures in Table 2 reflect some general differences between the WWW Health corpus and the other WWW domain corpora? Secondly, would the figures of syntactic complexity recorded from the WWW Health corpus change if the lemmas were encoded on the basis of another corpus consisting of other, edited and controlled text types?

The first question will be answered very briefly here. The texts from the administration, finance and commerce domains were originally produced for the printed medium such as reports, information material from authorities (cf. Jørgensen et al, 2002), characterised by explicit and precise formulation. The health domain texts were mainly intended for electronic communication, they reflect low-level information formulated in a style similar to informal, spoken language. Obviously, there is a close relationship between the degree of detail and precision of the content and the linguistic formulation. The second question will be answered in section 2.6.

2.3 The WWW Health Corpus

One of the domains is the health/medicine domain. The items of the OS are used to identify relevant texts which contain at least one of the items in the OS. Because of the hierarchical organisation of the labels (or items), it is possible to select the level of abstraction of the search items, e.g. the label *sygdomsbehandling* (disease treatment) covers texts concerned

with *medicinsk* (medical), *klinisk* (clinical), *kirurgisk* (surgical), etc. types of treatment, and texts describing *kirurgisk behandling* can deal with various subtypes of surgical intervention, e.g. *by-pass, kosmetisk operation, transplantation* (bypass, cosmetic operation, transplant surgery), etc. As to the content, by selecting an appropriate set of search items, an arbitrary composition of the corpus was avoided. The texts downloaded are edited into a corpus, henceforth referred to as the WWW Health Corpus.

2.4 The SDE Health Corpus: Texts from the Danish National Encyclopedia (SDE)

This corpus is a body of machine-readable health/medical articles written for the Danish National Encyclopedia (*Den Store Danske Encyklopædi*, henceforth SDE). It differs in several respects from the previously discussed one. The most prevalent features of the SDE corpus are the following. Its size is only about half of that of the web-based corpus. The texts are intended for the print medium with a clear, uniform explanatory aim.

FEATURES	WWW Health Corpus	SDE Health Corpus
Subject coverage	Broader domain coverage	Narrower, more specialised
Original Text medium	Online information and communication at various health/medicine sites	Machine-readable version of material originally published in print only
Topics in focus	*Health* (hospital and health service, health care, nursery, patient treatment, nutrition, preventive and alternative medicine ...)	*Medicine* (disease diagnostics, treatments, health care, preventive med., nursery, nutrition, pharmacy...)
Corpus size	~1.2 M tokens/35,000 lemmas	~0.5 M tokens/32,000 lemmas
Text types	Miscellaneous: public information, medical records, case reports, patient/doctor communication lines (FAQs /answers)	Homogeneous: articles from the Danish National Encyclopedia only
Communication level (Source to Target)	Authors: varying from domain experts to laymen Readership: web users, unspecified	Authors: Domain experts Readership: Laymen with a certain level of educational skills
Average level of formulation	Varying from formal to 'relaxed', often very close to spoken language	Homogeneous academic, 'polished'
Comments on the material	Only a limited number of the texts are edited and/or controlled	Content and language/of each article is controlled by experts and editors

Table 2: Characteristics of the two health corpora (Overview)

2.5 Basis for investigations

The common domain content provides a basis for the comparability of the two corpora. For the reasons outlined above, the SDE taxonomy for the health domain is used as the frame of reference. This ensures that the present investigation operates on a topic intersection of a certain size in the corpora.

In order to answer the second question raised in section 2.2, we investigated some prevalent tendencies in language use on the web and in encyclopedia text, of which the prepositional complementation of nouns is discussed in this paper. We are well aware of the fact that the limited size and the coverage-based composition of the corpus do not allow for general conclusions about the web as such being an appropriate resource for lexicography as discussed by Grefenstette (2002).

In STO, the syntactic structures of a lemma are described in terms of subcategorisation, based on principles of the valence theory (Helbig and Schenkel, 1969:12 ff.). A syntactic pattern (formalised in a *syntactic description*) contains first the information about the number of complements. For each complement then, its syntactic function (subject, object, etc.), syntactic form and category (valency-bound preposition, phrase and/or sentence type) are provided; and the complements are marked with regard to obligatoriness (yes/no). The STO lexicon model and basic description methods are discussed e.g. in Olsen (2002).

The examination of corpus instances dealt with the identification of arguments and a subtype of modifiers realised as prepositional complements for the following reasons. Firstly, an inspection of corpus evidence of selected words revealed some elements in their immediate context which traditionally would be categorised as adjuncts, being almost as frequent as complements. These elements are termed by Somers (1987) 'Middles': they make up borderline cases between arguments and modifiers. Secondly, an examination showed some significant differences in the two health corpora as regards the number and selection of possible arguments realised for the same word, i.e. the complements.

In our decision concerning the question whether middles should be registered in the syntactic description or not, we are inspired by the discussion in Somers (op.cit.). In addition to obligatory and optional complements to a certain extent also middles are encoded as subcategorised in STO. In natural language processing, one of the most crucial tasks is the proper recognition and production of syntactic structures, esp. the attachment of prepositional phrases. STO fulfils these requirements by providing very detailed descriptions of syntactic features, and in relevant cases including both the valency-bound complements and middles occurring frequently in the particular corpus.

2.6 Language investigations on the two health corpora

For the purpose of examining the realisations of arguments, presence/salience and phrase type of middles in our corpora, a list of 200 domain words were set up. The list comprised a variety of lemma types (nouns, verbs and adjectives) – all being frequent in both corpora and having a syntactic construction potential of a certain complexity. The concordance of the search lemma were analysed according to the criteria of complement and middle representations, their phrase types and frequency. The investigation presented here deals

only with nouns, which is the largest category of the Danish vocabulary. An extract of the search word list for nouns is shown below.

Nouns

Simple and compound nouns denoting aids, instruments, symptoms, diseases (65) (e.g.: *protese* 'prosthesis', *anfald* 'attack', *smerte* 'pain', *betændelse* 'inflammation', *dosis,* 'dose')

Deverbal result and/or process nouns, partly preserving the verbal subcategorisation properties (80) (e.g.: *operation* 'operation', *behandling* 'treatment', *forebyggelse* 'prevention', *undersøgelse* 'examination', *vaccination* 'immunization, *transplantation* 'grafting', '*indsprøjtning* 'injection')

2.7 An illustrative example

Table 3 exemplifies the findings in both domain corpora, summarised for the search word *operation* ('surgical intervention'). For reference purposes, the parallel figures in the Berlingske newspaper corpus of 20 M tokens (general language) are also shown. The field and figures in grey relate to a meaning of the search word external to the health domain, occurring only in the Berlingske newspaper corpus viz. 'military action'.

For each corpus, the following figures are provided: the number of lemma occurrences in total and with a relevant prepositional phrase only (and the ratio in %). Prepositional phrases expressing local, temporal and modal adverbial modifications are not counted as complements or middles.

Ex. No.	Lemma: *operation* + *preposition*	Semantic types of the complement	SDE Total: 506 with PP:96 (~19%)	WWW Total: 154 with PP:10 (~6.5%)	Berlingske Total: 807 with PP:28+6 (~4.2%)
1	*af* 'of'	Patient; illness	2; 7	0; 0	1; 2
2	*for* 'for'	Illness, reason	36	5	0
3	*gennem* 'through'	Organ	2	0	0
4	*hos* 'on'	Patient	3	0	1
5	*i* 'in'	Organ, body part	11	1	3
6	*med* 'with'	Method; instrument	18; 4	0	2;0
7	*på* 'on'	Patient; Organ, body part	2; 7	2; 2	0
8	*ved* 'at'	Illness	2	0	0
9	*ved_hjælp_af* 'by means of'	Method, instrument	2	0	0
10	*mod* 'against'	Enemy			6

Table 3: *Operation*: Prepositional complements and middles in three corpora

Instances selected from the SDE corpus

1 [...] **<operation> af** *små børn (small children)* / **<operation> af** *et åbent brok (an open rupture)*

2 [...] kirurgisk behandling, fx **<operation> for** *tandkødsbetændelse, (gum inflammation)*

3 Den kan [...] fjernes helt, oftest ved en **<operation> gennem** *det indre øre (inner ear)*

4 Dette kan gøres ved planlagt **<operation> hos** *[...] sunde personer (healthy persons)*

5 Dette muliggjorde udførelse af **<operationer> i** *brysthulen (thoracic cavity)*

6 [...] **<operation> med** *afmejsling af knoglefortykkelsen (chiselling off of the bone)*
[...] **<operation> med** [..]*meget fine instrumenter (very fine instruments)*

7 [...] **<operationer> på** *børn (children)* / **<operationer> på** *hjerte, lunger,* [...] *(heart, lungs).*

8 Der udføres også høreforbedrende **<operationer> ved** *otosklerose (othosclerosis)*

9 [...] gynækologiske **<operationer> ved_hjælp_af** *laparoskop (laparoscope)*

10 [...] styrker havde indledt en <operation> **mod** *armenske terrorister (Armenian terrorists)*

(7+2) [...] **<operation> på** *skjoldbruskkirtlen* (thyroid gland) **for** *forhøjet stofskifte* ('increased metabolism')

The figures of the syntactic complexity in the SDE Corpus are close the averages of the other domains, cf. Table 1. (The current figures for this corpus are: Noun/Reading ratio 1.07; with valence ~ 40%). The relationship between the text-type (and quality) and language complexity is clearly reflected by the results of the comparison. These observations have the following effect on the revision of the vocabulary encoded from the health domain. Firstly, they support the inclusion of prepositional phrases being salient or frequent middles specific to the domain (e.g. examples 3, 5 and 8). Secondly, they justify the method of accumulating the complements that potentially can co-occur, but in practice only rarely do so (e.g. a combination of 7 + 2) in one single syntactic pattern (cf. Olsen, 2002).

3 Summing Up

The outcome of the comparative examination of different corpora led us to the following conclusion. Firstly, the comparison of figures reflecting the differences between our www-based corpora showed that the syntactic complexity was extremely low in the WWW Health Corpus. Secondly, we investigated in more detail the question whether syntactic patterns of health domain nouns in general show less complexity than other domain nouns by using a control corpus. This investigation proved that although the vocabulary covered by the two corpora is quite similar, the syntactic structures of SDE Corpus are in general more complex than those observed in the WWW Health Corpus.

In conclusion, the use of a small-sized but high-quality corpus yields an essentially more convincing and varied basis for the encoding of syntactic information. Thus, the www-based corpus is not fully 'healthy' or appropriate to the needs of the STO project. In order to compensate for the deficiencies of the WWW Health Corpus a considerable augmenting as regards the text-types assembled is needed.

Acknowledgements

The STO lexicon project was granted by Danish Ministry for Science, Technology and Development. The list of project members and various presentations can be found at URL: http://cst.dk/sto/uk. The onomasiological structures were developed and applied by the staff at Copenhagen Business School. The texts of the Danish National Encyclopedia (SDE) were kindly made available for research purposes by Gyldendal Publishers, Copenhagen.

References

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. ELRA: Las Palmas.

Braasch, A. and Pedersen, B. S. 2002: *Recent Work in the Danish Computational Project "STO".* In: Braasch & Povlsen (eds.) Proceedings of the Tenth EURALEX International Congress, CST: Copenhagen.

Drost, J. 2002. *Sundhed. Onomasiologisk struktur.* (Unpublished, intern project document). CBS: Copenhagen.

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. EURALEX 2002.

Helbig, G. und Schenkel, W. 1969. *Wörterbuch zur Valenz und Distribution deutscher Verben.* Leipzig:VEB BibliographischesInstitut

Jørgensen, S.W., Hansen, C., Drost, J., Haltrup, D., Braasch, A., Olsen, S. 2003. *Domain Specific Corpus Building and Lemma Selection in a Computational Lexicon.* In: Corpus Linguistics 2003 Proceedings. Lancaster.

Olsen, S. 2002: *Some Aspects of the Syntactic Encoding of Nouns in a Computational Lexicon – the STO Project.* In: Braasch & Povlsen (eds.) Proceedings of the Tenth EURALEX International Congress. CST: Copenhagen.

Somers, H.L. 1987. *Valency and Case in Computational Linguistics.* EUP: Edinburgh.