STO: A Danish Lexicon Resource - Ready for Applications

Anna Braasch & Sussi Olsen

Center for Language Technology
Njalsgade 80
DK-2300 Copenhagen S, Denmark
anna@cst.dk, sussi@cst.dk

Abstract

This paper deals with the STO lexicon, the most comprehensive computational lexicon of Danish developed for NLP/HLT applications, which is now ready for use. Danish was one of the 12 EU-languages participating in the LE-PAROLE and SIMPLE projects; therefore it was obvious to continue this work building on our experience obtained from these projects. The material for Danish produced within these projects – further enriched with language-specific information - is incorporated into the STO lexicon. First, we describe the main characteristics of the lexical coverage and linguistic content of the STO lexicon; second, we present some recent uses and point to some prospective exploitations of the material. Finally, we outline an internet-based user interface, which allows for browsing through the complex information content of the STO lexical database and some other selected WRL's for Danish.

1. Project Objectives and Background

The objective of the Danish STO project (SprogTeknologisk Ordbase, i.e. Lexical Database for Language Technology) was two-fold: first, to develop a flexible, large-scale lexical resource in order to remedy a general bottleneck problem for Danish NLP applications; second, to strengthen the position of Danish, as a member of the still growing multilingual NLP/HLT community.

The project background as well as various development aspects and stages were presented at previous LREC Conferences (Braasch et al. 1998, Braasch & Olsen 2000, Braasch 2002, Olsen 2002)[1]. STO was a national collaborational project, initiated by CST and founded on a contract with the Danish Ministry for Science, Technology and Development. The duration of the project was three years, ending February 2004.

STO is well integrated with the European activities in the field of computational lexicon development for the following reasons. Danish was one of the 12 EU-languages that were part of the PAROLE (LE2-4017, 1996-98) and SIMPLE (LE4-8346, 1999-2000) projects. The LE-PAROLE/SIMPLE[2] models and descriptive methods obtained a status of being 'de facto standards' in the development of computational lexicons (Lenci et al, 2000). It was obvious to continue this work and build on our experience gathered from these multilingual projects, the more so as other national projects, e.g. the Italian CLIPS (Ruimy et al, 2002) were started on the same basis.

Even though in STO a number of language-specific refinements, various adaptations and extensions are implemented, its model and descriptive method is kept compatible with the architecture and descriptive language shared by the lexicons developed within the PAROLE/SIMPLE framework. Therefore the STO lexicon can be linked to other lexicons that share the same features and can be exploited in the development of multilingual lexical resources.

It was equally important to ensure that the descriptions of language-specific phenomena were compatible with Danish linguistic tradition and lexicographical practice. To this end, relevant bodies like the Danish Language Council and a number of field experts were consulted during the project in case of doubt. In preparation for reuse and data exchange in a monolingual environment, the information content of STO is also kept compatible with the national Danish Standard (1998) for lexical data collections, which comprises computational lexicons.

2. State of the Art, End of February 2004

2.1 Lexical Coverage

Table 1 shows the actual number of lemmas in the database, and to which extent the different word classes have been provided with a) only morphological information, b) with morphological and syntactic information c) with morphological, syntactic and semantic information.

Lexical Category	Lemmas No.	Morph. only	Morph. & Synt.	Morph. & Synt. &Sem.
Noun	**64740**	47%	41%	12%
Adjective	**9770**	32%	55%	13%
Verb	**5775**	2%	81%	17%
Adverb	**770**	81%	19%	
Interjection	**160**	100%	0%	
Preposition	**80**	100%	0%	
Conjunction	**60**	100%	0%	
Pronoun	**45**	100%	0%	
Misc.	**130**	100%	0%	
Total	**81530**			

Table 1: The contents of the STO database

The contractual commitments were the following: A lexical database of 45,000 lemmas in total, divided between

- Approx. 30,000 from general language
- Approx. 15,000 from six different domains of language for special purposes
- All lemmas to be provided with detailed morphological and syntactic information
- Detailed semantic information on a subset of general language lemmas

[1] These presentations can be downloaded from http://cst.dk/sto/uk

[2] http://www.ub.es/gilcub/SIMPLE

- Ontological information and selectional restrictions on the lemmas of one domain.

As it can be seen, the total figures of words with morphological information in the database by far exceed the number of words originally planned. This is due to the fact that the semi-automatic morphological encoding procedure which we used proved very efficient. (The large number of lemmas only with morphological information considerably increases the usability of the lexicon in applications such as shallow parsers, taggers, etc.)
Only the contractual number of words has been encoded with syntactic and semantic information. This goes for nouns, verbs, adjectives and adverbs; the closed word classes have not been provided with syntax and semantics. The words for the syntactic encoding were selected on a frequency basis in order to reach the required amount, which means that all verbs are provided with syntax, whereas only nouns and adjectives with a frequency higher than 20 are provided with syntactic information.
As regards the semantic information, approx. 7000 lemmas from general language have been provided with detailed semantic information (level 3 semantics, cf. 2.3.1. below) and 2500 from a specific domain with more restricted semantics (level 2, cf. 2.3.1. below). All words in STO have reference to the source domain (level 1).

2.2. Domain Vocabulary

The domain vocabulary is selected from corpora created by collecting texts from the web. This approach was chosen since texts on the web are easily available and of a kind that are suitable for the vocabulary that STO is supposed to cover, i.e. not highly specialized terms but the vocabulary that laymen might encounter in various contexts. See Jørgensen (2003) and Olsen (2002) for further information on the collection of texts and the selection of lemmas. The domains and the distribution of lemmas on word classes for each domain are shown in table 2.

Domain	Nouns	Verbs	Adjectives	Total
IT	1730	160	115	2005
Environment	1770	50	300	2120
Commerce	1800	60	160	2020
Administration	2430	25	220	2675
Health	2285	40	250	2575
Finance	1880	30	160	2070
Total				**13465**

Table 2: Domain vocabulary and part of speech distribution

All lemmas from the domains have been encoded with both morphological and syntactic information. The health domain has furthermore been encoded with semantic information of specificity level 2 (cf. below).
Based on preliminary studies, each domain was planned to include about 2500 lemmas selected from corpora of between one and two million tokens each. The final number of lemmas is, however, slightly lower. To verify whether the current domain vocabularies are adequate is not a simple task. A comparison with existing lemma lists is difficult, firstly because the delimitation of a domain varies a lot, secondly because lemma lists are made for different purposes and address different users, and thirdly because there are very few Danish lemma lists available for the different domains. Nevertheless an experimental comparison with an existing lemma list for the environment domain showed that the STO vocabulary for this domain was indeed adequate.

2.3 Linguistic Coverage

The organisation of the linguistic information of STO is based on the conceptual and representational model of PAROLE/SIMPLE. A lemma is described by a combination of morphological, syntactic and semantic units; each unit represents a particular linguistic behaviour of the lemma (cf. Braasch, 2002, Braasch et al, 2004).

2.3.1 General Linguistic Features

The general linguistic features described at the three layers are the following:

- Morphology: PoS, inflectional patterns, agreement features, compounding properties, etc.
- Syntax: subcategorisation frame (comprising categorical and functional valency), diathesis and alternation phenomena, reflexivity of verbs, etc.
- Semantics: the information is provided at three specificity levels. Level 1 contains domain reference only. Level 2 comprises domain information, ontological type, argument structure and selectional restrictions, whereas level 3 is identical with the SIMPLE semantics (cf. Lenci et al, 2000; Pedersen & Paggio, in press). Information types of level 3 are ontological type, semantic relation, argument structure, selectional restrictions, qualia structure, event structure, domain information, etc.

The subdivision of the semantic information into three levels is introduced for practical reasons. Level 2 and level 1 are proper subsets of level 3 representing a relatively lean semantics. At a longer term these lean levels can be extended to level 3 descriptions.

2.3.2. Language-specific Extensions

The extension of a lexical resource produced in a multilingual environment into a large-scale, monolingual resource requires a number of modifications. In particular, advanced monolingual applications benefit by detailed language-specific information. To this end, we implemented the following enhancements.

- Adverbs:

The granularity of syntactic descriptions is increased considerably. Their description in the STO lexicon differs from earlier ones by the individual and corpus based coding of each adverb with respect to position, type of head and a set of other syntactic characteristics and also by making a clear distinction between semantics and syntax, since all the properties described can be tested syntactically, see further in Nimb, 2004).

- Nouns:

At the morphology layer, information on compounding is implemented, at the syntactic layer the definition of syntactic frame elements is reconsidered, viz. ‘middles’ as defined in Somers (1987) are included on the basis of significant frequency in the corpus.

- Adjectives:

The description at the syntactic layer now distinguishes between different types of optionality (viz. complements being both syntactically and semantically optional, versus

those being syntactically optional but semantically obligatory). Encodings of optionality which might cause over-generation are especially marked.

- Verbs:

For this part of speech only minor revisions have been made.

3. Validation

In a collaborative lexicon project like STO, it is a key issue to ensure the inter-coder consistency in order to achieve homogeneity of the linguistic content. To this end, the lexicographers were guided by detailed encoding guidelines and worked with encoding tools supporting consistency checks. The successive stages of the work were organized in three steps: The lemmas were encoded by one lexicographer/team and then checked/revised by another. Finally, all data were validated at CST before uploading into the STO database. Also external users' reported experience and relevant comments were taken into consideration during the process.

4. STO Data in User Applications

Because of the fact that STO is currently the largest and most comprehensive computational lexicon for Danish, a growing demand for this resource is taken for granted. STO material is already being used in a number of projects and applications, for a variety of purposes. According to users' specifications, data subsets were extracted from the lexicon. These were adapted to various format requirements and the linguistic content was exploited for both particular research and development purposes. This way both the linguistic content and the formal properties of the lexicon were judged from the user's points of view. A few examples below illustrate some typical uses of STO-data.

In research, data were exploited e.g. for

- Evaluation of search engine behaviour in a multi-lingual environment (Data subset: inflectional morphology of nouns; German research project).
- Computational analysis and processing of complex sentence structures from the point of view of potential reading speed (Data subset: verbs with selected subcategorisation frame types; PhD. thesis).
- Conversion of verb entries into the lexicon format of the Danish Dependency Treebank (Data subset: approx. 2.000 verbs with comprehensive syntactic descriptions; linguistic research project).
- Using the qualia structure information for the calculation of semantic relations in compounds (Data subset: nouns with SIMPLE semantics, research).

In practical applications, data were used e.g. in

- Machine translation for a specific domain. (Data delivered: about 3,000 lemmas with morphological information, encoding of new lemmas using the STO encoding tools and methods.)
- Lemmatiser for Danish. (Data delivered: approx. 70,000 lemmas with morphological information for the software development.)
- Information retrieval system. (Data delivered: 700 lemmas with morphological and semantic information for content-based querying (OntoQuery).)
- Preparatory work with the aim of exploitation of verb descriptions in a construction dictionary for humans. (Data delivered: selected verbs with syntactic descriptions for a test application at a dictionary-publishing house.)
- Ongoing development for speech technology applications. (Data delivered: all lemmas with all inflected forms and syntactic descriptions for enrichment with phonetic transcription.)
- Ongoing negotiations concerning development of e-learning software. (Data to be exploited: a selected subset of the lemmas provided with morphological and syntactic information in computer-aided language learning: Danish as second language.)

Reports on successful experimental applications and positive responses from the users provide a promising basis for the marketing of the STO resource both for the research community and for commercial NLP/HLT tool developers.

5. Future Developments of the Material

The main goals for further developments are the following

- Increasing the number of encoded semantic senses considerably (up to 80,000).
- Addition of frequent collocation types provided with information of the three descriptive layers (some experimental work carried out on this task is described in Braasch & Olsen, 2000).
- Enlargement of the lexical coverage up to 200,000 entries, all provided with morphological and syntactic information.
- Enhancement of the linguistic descriptions by adding pronunciation (in form of transcription and/or sound).

The labour-intensive processes in these developments will be supported by statistically based, automatic and semi-automatic tools - together with the web as a dynamic corpus. Human inspection will still vouch for the satisfaction of the quality requirements.

5.1. Perspectives for Exploitation of the Data in Further Applications

Currently, only few industrial products are developed for Danish at all, partly due to the bottleneck-problem of a lacking lexical resource. Because of its comprehensive information content, STO can keep up with very different demands and it can be exploited as a lexicon component in both monolingual tools (parsers, taggers, authoring tools, browsers, spelling/grammar checkers) and in multi-lingual applications (MT systems, search engines, etc.).

On account of the fact that the lexical information types and the data structure of STO are conformant with the reference computational model used in specifying the various components of MILE[3] (e.g. in Atkins et al., 2002), the STO lexicon could be appropriately integrated into the multilingual MILE architecture as one of the monolingual modules. The enrichment of STO with more comprehensive lexical semantic descriptions will provide the necessary basis for correspondence links.

[3] MILE stands for Multilingual ISLE Lexical Entry (ISLE: International Standards for Language Engineering)

6. User Interface

In addition to various NLP applications, STO is a valuable resource also to linguistic researchers, teachers and learners of the Danish language as seen in section 4. In order to provide easy access to the material, a web interface was developed to the database, facilitating word searches, searches for syntactic patterns etc., and corpus investigations. The interface is accessible at URL: **http://cst.dk/sto/uk**.

6.1. Search options

- Word Search displays all inflected forms and syntactic constructions of the lemma.
- Search of compounds displays all compounds containing the search lemma as one of its elements.
- Corpus Search: from each result of a Word Search links are established for direct searches in corpora. (Corpus instances are displayed in KWIC format.)
- Parameterized Search using a combination of a Part of speech and value(s) of all prevalent properties of the POS selected. (Up to 30 lemmas meeting the combination of search parameters are displayed.)

6.2. Additional facilities

The user interface provides links to other on-line WLR's of Danish, such as electronic dictionaries for human use (*Retskrivningsordbogen,* The Official Spelling Dictionary, and *NetOrdbogen,* The Internet Dictionary) and corpuses (*Korpus2000* and *Berlingske Tidende*, a newspaper corpus). Further, web sites in Danish can be accessed too using the Google search engine. These facilities allow users to search in additional resources e.g. for comparison and supplementing purposes in a user-friendly way.

7. Summing Up

In this paper we presented the Danish STO lexical resource being ready for different applications, and we outlined some tasks for its further development. One of the main future objectives is to move STO forward into a multilingual lexicon environment by enriching the existing material with the required semantic descriptions. Through the establishment of bi- and multilingual correspondences to other comparable monolingual lexical resources, Danish can be part of new technologies, such as internet-based translation solutions and large-scale localization services.

Acknowledgements

The project was carried out by a large number of staff members from the following institutions:

- Copenhagen Business School, Institute for Computational Linguistics.
- University of Copenhagen, Institute for General and Applied Linguistics,. University of Southern Denmark, Institute of Business Communication and Information Science.
- University of Copenhagen, Center for Language Technology (coordinator of the project).

References

Atkins, S., N. Bel, F. Bertagna, P. Bouillon, N. Calzolari, C. Fellbaum, R. Grishman, A. Lenci, C. MacLeod, M. Palmer, G. Thurmair, M. Villegas, A. Zampolli (2002). From Resources to Applications. Designing the Multilingual ISLE Lexical Entry. In: *Proceedings of the 3rd International Conference on Language Resources and Evaluation* (pp. 687-693). Las Palmas.

Braasch, A., C. Navarretta, S. Nimb, S. Olsen, B. Pedersen, C. Povlsen (2004). *Lingvistiske Specifikationer for STO*. URL: http://cst.dk/sto

Braasch, A. (2002). Current Developments of STO – the Danish Lexicon Project for NLP and HLT Applications. In: *Proceedings of the 3rd International Conference on Language Resources and Evaluation* (pp. 986-992). Las Palmas.

Braasch, A. & B. S. Pedersen (2002). Recent Work in the Danish Computational Lexicon Project „STO". In: *Proceedings from the Tenth Euralex International Congress* (pp.301-314), CST, Copenhagen.

Braasch, A. & S. Olsen (2000). Towards a Strategy for a Representation of Collocations – Extending the Danish PAROLE-lexicon. In: *Proceedings of the Second International Conference on Language Resources and Evaluation* (pp. 1009-1016), Athens.

Braasch, A., A. B. Christensen, S. Olsen & B.S. Pedersen, (1998). A Large-Scale Lexicon for Danish in the Information Society. In: *Proceedings of the First International Conference on Language Resources & Evaluation* (pp. 249-254). Granada.

Danish Standard (1998). DS2941-1. Leksikalske datasamlinger. Indholds- og strukturbeskrivelse. Del 1. Dansk Standard. Copenhagen.

Jørgensen, S. W., C. Hansen, J. Drost, D. Haltrup, A. Braasch, S. Olsen (2003): Domain specific corpus building and lemma selection in a computational lexicon. In: *Corpus Linguistics 2003 Proceedings* (pp. 374-383). Lancaster.

Lenci, A., N. Bel, F. Busa, N. Calzolari, E. Gola, M. Monachini, A. Ogonowski, I. Peters, W. Peters, N. Ruini, M. Villegas, A. Zampolli (2000). SIMPLE: A General Framework for the Development of Multilingual Lexicons. In: *International Journal of Lexicography,* Vol. 13, no. 4, (pp.249-263). OUP.

Nimb, S. (2004). A corpus-based syntactic lexicon for adverbs. In: *Proceedings of the 4th International Conference on Language Resources and Evaluation.* Lisbon.

Olsen, S. (2002). Lemma selection in domain specific computational lexica – Some specific problems. In: *Proceedings of the 3rd International Conference on Language Resources and Evaluation* (pp.1904-1908). Las Palmas.

LE-PAROLE, (1998). *Danish Lexicon Documentation.* Internal report, Center for Sprogteknologi, Copenhagen.

Pedersen, B., Paggio, P. (In Press) The Danish SIMPLE Lexicon and its Application in Content-based Querying. To appear in *Nordic Journal of Linguistics*. University of Liège.

Ruimy, N., M. Monachini, R. Distante, E. Guazzini, S. Molino, M. Ulivieri, N. Calzolari, A. Zampolli (2002). Clips, a Multi-level Italian Computational Lexicon: a Glimpse to Data. In: *Proceedings of the 3rd International Conference on Language Resources and Evaluation* (pp. 792 – 799). Las Palmas.

Somers, H. L. (1987). *Valency and Case in Computational Linguistics*. Edinburgh University Press.