A Large-Scale Lexical Database of Danish for Language Technology Applications and Other Purposes

Anna Braasch

Anna Braasch was born in Budapest in 1946, studied languages at ELTE University, and has a Master in German from both Gothenburg and Copenhagen universities with Minors in Computational Linguistics and Russian. She is a senior researcher at the Centre for Language Technology (CST), University of Copenhagen, and managed the STO project. Her expertise is in computational lexicography and terminography, her main research interests are phraseology and lexical semantics, and her language skills include Danish, English, German, Hungarian, Russian and Swedish. She was on the boards of the Danish and Nordic lexicographer associations (respectively LEDA and NFL), was President of NFL, and is currently the Vice-President of Euralex.
anna@cst.dk

1. Background

SprogTeknologisk Ordbase (STO, Lexicon for Language Technology Applications) is the most comprehensive computational lexicon of Danish, primarily developed for Natural Language Processing (NLP), including commercial language technology products and computational linguistic research purposes. STO is created within the framework of a national project, led by the Centre for Language Technology (Center for Sprogteknologi, CST) at the University of Copenhagen. The work was carried out as a result of a contract with the Danish Ministry for Science, Technology and Development. The three-year project ended in February 2004.

CST initiated the project and was responsible for its management and the coordination of the work, including tasks such as software development, the creation of linguistic specifications and definition of guidelines for encoding.

The lexicon material was produced in cooperation with the Institute for Computational Linguistics of the Copenhagen Business School, the Institute for General and Applied Linguistics of the University of Copenhagen and the Department of Business Communication and Information Science of the University of Southern Denmark.

The project members had various skills relevant to the task comprising theoretical linguistics, terminology, lexicology, corpus linguistics, computational lexicography and linguistics, and database knowledge. This wide range of expertise ensured that the lexicographic work carried out was of high quality, which is reflected in the lexicon.

The need for a computational lexicon

The main objective of the STO project was the development of a computational lexicon of Danish for a broad practical application area. Language industry and research into computational linguistics often experience the lack of a large-scale comprehensive lexicon as a bottleneck problem for the development of most applications and software tools for language engineering. In particular, for less widely-spoken languages such as Danish, it is essential to develop some multipurpose and flexible language resources in order to optimize the cost/benefit ratio.

In this sense, the STO serves as a basic lexical data collection from which various dedicated modules can be derived for particular applications, such as lemmatizers, inflection analyzer/generators, shallow parsers or Danish modules for Machine Translation (MT) systems.

2. The lexicon

The most important features of the STO lexicon are, besides its size, being theoretically well-founded and empirically supported. The descriptions are very detailed, each piece of information is labelled explicitly and precisely, and any item is easily accessible as the entire lexicon is structured and stored in a relational database (ORACLE). Thus, it is straightforward to extract, for example, all syntactic frames of a lemma, or all lemmas sharing the same syntactic frames or a particular syntactic construction, for research purposes.

The composition of the lexicon

The STO lexicon contains over 81,000

Lexical Category	No. of Lemmas	Morphology only	Morphology & Syntax	Morphology & Syntax & Semantics
Noun	64735	47%	41%	12%
Adjective	9773	32	55%	13%
Verb	5775	2%	81%	17%
Adverb	771	81%	19%	
Interjection	158	100%		
Preposition	80	100%		
Conjunction	60	100%		
Pronoun	44	100%		
Misc.	128	100%		
Total	81524			

Table 1: The composition of the entire STO vocabulary

lemmas, of which approximately 14,000 come from six different domains of language for specific purposes (LSP). All lemmas are provided with lexical category information and exhaustive descriptions of their inflectional properties, and 45,000 of them with a fine-grained syntactic description as well. Tables 1 through 3 show the composition of the vocabulary covered in detail. The STO database is not intended to cover highly specialized terms, but focuses on words of the domain languages that laymen will have to read and understand as part of everyday life. This is considered a kind of transitional area between the general language and specialized expert languages.

General language and domain language vocabulary

Table 1 shows the composition of the entire STO vocabulary classified by the feature 'Lexical Category' (in other terms: 'word class' or 'part of speech') and to which extent the different word classes have been provided with either (a) only morphological information, or (b) morphological and syntactic information, or (c) morphological, syntactic and semantic information.

The large number of lemmas with only morphological information is especially useful in applications such as shallow parsers, taggers, spell checkers, etc.

The words for the syntactic encoding were selected on a frequency basis; all verbs are provided with syntax, whereas only nouns and adjectives above a certain frequency threshold are provided with syntactic information.

Tables 2 and 3 specify the vocabulary from the selection point of view (as originating from general language and domain language texts).

Lexicon model

The establishment of the descriptive model and the linguistic specifications for STO greatly benefit from the experience acquired at CST within the framework of the multilingual (LE2-4017) PAROLE project (1996-98) of the European Commission. The PAROLE lexicons were built around a generic model, an instantiation of the EAGLES recommendations in an enriched GENELEX model (details about the EAGLES and GENELEX projects are available as part of the PAROLE information: http://hltcentral.org/usr_docs/project-source/parole/ParoleFinal.pdf). Thus, the Danish STO lexicon is well integrated in the multilingual infrastructure of European computational language resources, which ensures its compatibility with other resources developed for Human Language Technology (HLT).

Linguistic description

The STO lexicon is corpus-based as regards both the selection and the description of lemmas. The linguistic descriptions are based on corpus analysis, and all lemma types are treated in a uniform way.

The linguistic information content of the STO lexicon is organized according to the traditional practice in computational linguistics of division into three independent descriptive layers, i.e. morphological, syntactic and semantic. Each descriptive layer is made up of a comprehensive system of the characteristic linguistic properties. The linguistic description of a lemma is structured in different sets of information, the so-called units; each unit represents a particular morphological, syntactic or semantic behaviour of the lemma at the layer concerned. From the computational point of view, a unit is a structured object containing a feature-based description expressed in attribute/value pairs.

The general linguistic features described at the three layers are as follows:

- Morphology: lexical category, inflectional patterns, spelling variants, agreement features, compounding properties, etc.
- Syntax: syntactic patterns comprising subcategorisation frame (categorical

Center for Sprogteknologi (CST)

The CST (Centre for Language Technology) is a research institute at the University of Copenhagen. It employs a staff of approximately 20, including computational linguists, lexicographers, computer scientists and engineers. Its mandate is to carry out and promote strategic research and development in language technology (LT) and computational linguistics in Denmark. It addition to research activities, the CST has considerable experience in the development of a variety of LT applications for both national research projects and commercial purposes, including Machine Translation, XML mark up, company specific ontologies and inter-institutional term databases for EU institutions and agencies, as well as controlled language and authoring tools.
http://cst.dk

Lexical Category	Number of Lemmas
Noun	52840
Adjective	8568
Verb	5410
Adverb	771
Interjection	158
Preposition	80
Conjunction	60
Pronoun	44
Misc.	128
Total	**68059**

Table 2: General language vocabulary (all closed word classes belong to this category)

Domain	Nouns	Verbs	Adjectives	Total of Domain
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	11895	365	1205	13465

Table 3: Domain language vocabularies in the STO database with part of speech distribution

STO availability

The lexicon material is now available for both commercial use and research purposes. Starting this summer, the licensing and distribution will be handled by the Evaluations & Language resources Distribution Agency. (ELDA, http://elda.org).

Standard packages

- 81,000 entries with morphological description only, provided with full documentation of the morphological layer;
- 81,000 entries with morphological description whereof 45,000 entries are provided with syntactic information including full documentation of both layers.

Standard delivery formats

- Morphology in comma-separated plain text files
- Syntax in XML format
- ORACLE dump database files (8.1 on request)

User defined data package types and delivery formats can be produced on demand.

Additional facilities

The STO web interface provides links to other on-line Danish language resources, such as electronic dictionaries for human use (*Retskrivningsordbogen*, the Official Spelling Dictionary, and *NetOrdbogen*, the Internet Dictionary) and corpora (Korpus2000 and Berlingske Tidende, a newspaper corpus). In addition, Danish websites can be searched through a link to Google. These facilities enable direct searches in a user-friendly way, e.g. to compare the STO data with information in the electronic dictionaries and supplement the STO data by corpus evidence.

http://cst.dk/sto/uk

and functional valence), diathesis and alternation phenomena, reflexivity of verbs, etc.

- Semantics: the information is provided at three specificity levels. Level 1 contains domain reference only (all entries). Level 2 comprises domain information, ontological type, argument structure and selectional restrictions (about 2,000 entries). Level 3 is identical with the SIMPLE semantics. Information types of level 3 are the ontological type, semantic relation, argument structure, selectional restrictions, qualia structure, event structure, domain information, etc (about 7,000 entries). The subdivision of the semantic information into three levels is introduced for practical reasons. Levels 1 and 2 are proper subsets of level 3, representing a relatively lean semantics.

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, then checked/revised by another, finally all data were validated at CST before uploading it to the STO database. The reported experience and comments from external users were taken into consideration during the process.

3. STO data in use

STO is currently the largest and most comprehensive computational lexicon for the Danish language, and the demand for this resource is growing. The 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. The examples below illustrate some typical uses of the STO-data.

In research

- evaluation of search engine behaviour in a multilingual environment
- computational analysis and processing of complex sentence structures from the point of view of potential reading speed
- conversion of verb entries into the lexicon format of the Dependency Treebank
- testing of a computational grammar for Danish
- using the qualia structure information to calculate semantic relations in compounds

In practical applications

- MT for a specific domain
- lemmatizer for Danish
- information retrieval system prototype
- preparatory work with the aim of exploitation of verb descriptions in constructing a dictionary for humans
- ongoing development for speech technology applications, extension with pronunciation of all word forms

Perspectives for further exploitation

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

Currently, only few industrial products are developed for Danish at all, partly due to the bottleneck problem of lacking a lexical resource. Because of its comprehensive and detailed content, STO can keep up with very different demands and be exploited as a lexicon component in both monolingual tools (parsers, taggers, authoring tools, browsers, spelling/grammar checkers) and in multilingual applications (MT systems, search engines, etc) as well as for HLT tasks such as developing computer-aided language learning tools for Danish as a second language, etc.

4. User interface: looking up single entries

In addition to various NLP applications, STO offers a valuable resource to linguistic researchers, teachers and learners of the Danish language. To facilitate access, there is a web interface that enables various searches and corpus investigations. However, the database contains more linguistic information than shown on the screen for human users.

Search options

- Word Search displays all the inflected forms and syntactic constructions of the lemma.
- Compound Search displays all the compounds containing the search lemma as one of its elements.
- Corpus Search establishes links from each result of a Word Search to direct searches in corpora.
- Parameterized Search uses a combination of the lexical category and the value(s) of all its selected prevalent properties.