Current Developments of STO - the Danish Lexicon Project for NLP and HLT Applications

Anna Braasch

Center for Sprogteknologi
Njalsgade 80
DK-2300 Copenhagen S
anna@cst.dk

Abstract

The Centre for Language Technology (Center for Sprogteknologi, CST) is in charge of a national project developing a large-scale Danish lexicon for HLT and NLP applications. The short name of the project is *STO*, which stands for *S*prog*T*egnologisk *O*rdbase (Lexical Database for Language Technology). The project is inspired by principles and methods applied in the multilingual LE-PAROLE project (1996-98) the aim of which was to develop harmonised written language resources for 12 EU languages. The Danish PAROLE lexicon was produced by CST and the STO project highly benefits from the experience acquired from the work mentioned. This paper deals with a few central tasks of the ongoing project. It discusses the development of a smaller lexical resource produced in a multilingual environment into a large-scale, monolingual resource. Two different methods of increasing the vocabulary will be presented in detail; the extension of the linguistic coverage and the refinement of the linguistic description by including more detailed language-specific information. Finally, some exploitation perspectives and the development of an internet-based user-interface will be presented. The STO project gets funding from the Danish Ministry for Science, Technology and Development for a period of three years (2001-2004).

1. Introduction

The Danish STO project greatly benefits from the Danish lexicon material developed at CST within the multilingual LE-PAROLE project. (1996-98). In this sense, the groundwork for the STO lexicon was laid in the PAROLE project as regards the model, descriptive language and methodology of linguistic description.

The objective of the PAROLE project was to elaborate machine-readable text collections and lexicons for 12 languages sharing linguistic specification descriptive model and information structure. The expected outcome of this project was – besides the production of the resources mentioned - the initiation of new, national and co-operative 'follow-up' projects reusing and extending the material elaborated. The STO project has to be seen and understood within this context (Braasch et al., 1998).

2. Background

The Danish PAROLE lexicon contained approx. 20,000 lemmas from general language supplied with a description of their morphological and syntactic properties – thereof approx. 8,500 were supplied with semantic information in the framework of the SIMPLE project. This material serves as a point of departure for the STO project.

3. Project objectives

The objective of the STO project is to develop a computational lexicon of Danish for a broad practical application area (human language technology and natural language processing). Language industry and research into computational linguistics often experience the lack of a large-scale comprehensive lexicon, which appears to be a bottleneck problem for most applications. In particular, for less widely spoken languages such as Danish it is essential to develop some multipurpose and flexible language resources in order to optimise the cost/benefit ratio. In order to satisfy as many application requirements as possible, the lexicon must contain a large amount and variety of information, in-depth linguistic descriptions – all stored in a structured and easily manageable way.

In this sense the STO lexicon will serve as a basic lexical data collection from which various dedicated lexicon modules can be derived. Pronunciation is an information type not yet included in the lexicon, although it is essential for speech technology. In planning new developments it will be taken into consideration and the mode of presentation – transcription an/or voice - will be discussed.

In order to develop a resource of a usable but also realistic size the principal goal of the STO project is to grow the resource size from 20,000 lemmas to 50,000 as a first extension. This development is not merely concerned with the populating of the lexicon by simply adding more lemmas and descriptions, but it also includes a revision of various basic theoretical and practical issues. In what follows, we will discuss a few central issues from the computational and linguistic work carried out.

4. Descriptive language and information structure revised

The PAROLE linguistic specifications and the information structure were designed for sharing by 12 languages, which, of course, had to be revised for monolingual purposes. In this regard, two aspects of revision came into play in opposite directions.

The first one was the need for a significant reduction of the overwhelming number of linguistic features by eliminating those being irrelevant for Danish.

The second one was the enhancement of the linguistic specifications and, accordingly, the information structure in order to hold a number of additional language specific information types and their interrelations. The PAROLE data have been adapted to the STO requirements and the tailored entries have been integrated into the STO lexical database. The PAROLE data and structure were transferred into an ORACLE database and organised into a more user-friendly and intuitively understandable structure, from the lexicographer's, the customer's, and, last but not least, from the database manager's point of view.

However, it is important to remember that lexical databases of different languages produced on the basis of shared specifications like the 12 PAROLE lexicons should be kept compatible with each other in order to facilitate future bilingual or multilingual linking.

5. Developments within the linguistic area

This section deals with the extension of the lexical coverage (i.e. the number and origin of entry words) and of the linguistic coverage, i.e. the enhancement of linguistic description with language specific information types, addition of detailed features, treatment of new linguistic phenomena, etc.

5.1. Lexical coverage: the point of departure

The list of entry words in a PAROLE lexicon has been compiled on the basis of the following general requirements

?? Precondition: the number of entry words belonging to each part-of-speech was predefined

?? All closed word classes had to be fully covered

?? Existing computational lexicons had be reused as far it was feasible (lemmas and their descriptions).

The Danish lexicon had in addition the following properties:

?? Each lemma was also contained in the electronic version of the Official Danish Spelling Dictionary (Retskrivningsordbogen 1986) and occurred with a certain frequency in the first large-size Danish text corpus compiled by the Danish Dictionary (DDO) project.

?? Preferably, the lemma belonged to a frequent morphological pattern

5.2. Extending the lexical coverage: Selection of new lemmas for the STO lexicon

The first version of the STO lexicon is planned to contain approximately 50,000 lemmas fully described with morphological and syntactic information, hereof approx. 35,000 originating from general language (GL) lemmas and 15,000 from languages for specialised purposes (LSP) of six delimited domains, such as information technology, environment, etc. Figure 1 illustrates the composition of the STO vocabulary.

The methods of lemma selection used for general and specialised languages differ in several ways, notably because of the different resources used for these tasks. We sketch out the two methods below.

5.2.1. Selection of general language (GL) vocabulary

The basic steps of the GL lemma selection was to take the PAROLE lemma list as the point of departure (reuse of 17,500 lemmas) and select approx. 17,500 further lemmas. The selection of new lemma candidates was based on a list of general words we have been allowed to utilise. This list was produced within the framework of The Danish Dictionary Project (Den Danske Ordbog, DDO) for the work on a contemporary dictionary (to be published 2002/2003). The list was compiled automatically on the basis of existing dictionaries and corpus investigations. It contains approximately 100,000 words with frequency information based on a corpus of 40 million tokens. The frequency figures, however, contain some noise because corpus that was used was not tagged morphosyntactically. For the STO project we sorted out from this list the lemmas already encoded in PAROLE. A further manual selection during the encoding process discarded lemma candidates with low frequency figures or those that were perceived as dialectal, informal, old fashioned, etc. words and thus irrelevant or unsuitable for the coverage of STO. Also, erroneous frequency figures (being obviously too high or too low) are checked manually against the corpus in order to correct the ranking of the lemma in question.

5.2.2. Encoding the GL vocabulary

Although the encoding of GL lemmas is carried out manually, the process is heavily supported by computational methods and tools. One of the methods is to read and sort the lemmas backwards (i.e. to create a "reverse vocabulary"); in the resulting list all lemmas with the same ending appear together. In most cases, lemmas having the same ending (whether a derivative suffix or the last component of a compound) follow the same inflectional pattern. The encoding can therefore be done partly automatically, supported by human control.

The method of combining automatic and manual steps considerably speeds up the encoding of morphological information. At present, the number of GL lemmas provided with morphological description is approx. 45,000 and it is still growing fast. The cost/benefit ratio for this part of the encoding proved to be positive. Therefore, we estimate the number of lemmas provided with morphological information will be approximately 70,000 lemmas, although as planned, only 35,000 will be supplied also with syntactic descriptions. The advantage of more lemmas than planned provided with morphological information is obvious especially for language recognition applications, e.g. to be used as component in a morphological tagger or lemmatizer.

5.2.3. Selection of specialised language (LSP) vocabulary

As mentioned earlier, the STO lexicon will contain a vocabulary of approx. 15,000 lemmas originating from six domains of language. The domains are selected according their supposed relevance for language technology applications, namely IT, environment, commerce, health, public administration – one further domain is still to come. The lemmas selected should not be highly specialised terms of the domain, but rather words that laymen have to read and understand as part of their everyday life, words belonging to the so-called 'grey area' vocabulary. Thus, we deal with texts from expert to semi-expert or layman having a medium or low level of subject field competence.

This part of the lexicon extension is entirely based on work carried out in STO starting from scratch. For each of the domains mentioned we assemble a text corpus at least of 1 to 1,5 million tokens. We expect that from each corpus will be extracted 2,500 new LSP lemmas in average. Appropriate sources are less specialised documents such as textbooks, popular scientific magazines, newspapers, web publications, Users' manuals, etc. At the moment we have assembled two corpora (IT and environment), the vocabulary of the first domain is encoded at the morphological and syntactic layers.

The compilation of a list of lemma candidates involves several steps. The corpus size is in this phase of the project rather limited – because of the approach chosen: the main point is to compile smaller lists of lemmas for a variety of domains. The task of corpus-based selection of LSP vocabulary is discussed in (Olsen 2002).

5.2.4. Encoding the LSP vocabulary

The vocabulary of specialised languages often is source of difficulties requiring some special solutions. Firstly, it contains a large number of newly coined Danish words and foreign words and expressions without standardised spelling and/or inflection. Secondly, although there may exist a (usually bilingual) dictionary for the domain in question, it mostly does not provide other information about the lemma than a part-of-speech marking besides translation and/or explanation.

Specialized dictionaries are mainly concerned with encyclopedic information and the linguistic dimension is only sparsely represented (Bergenholtz & Tarp 1995). Therefore we often have to develop the linguistic description of novel and foreign lemmas from scratch.

In this process we - of course - highly rely on corpus evidence and frequency of forms as regards spelling and inflectional variants. However, this is not sufficient because of the limited corpus size: the problem of 'silence' in the list of forms i.e. a form assumed to be appropriate for the lemma is not or only very sparsely represented in the corpus. A closely related problem is the appearance of hapax legomena (Lebart et al. 1998)

In order to ensure the reliability of the morphological information encoded, the Danish Language Council is consulted being authorised to advise on both descriptive and normative aspects of spelling and inflection. Problems of hapax legomena are also discussed with domain experts, concerning their possible domain relevance and usage.

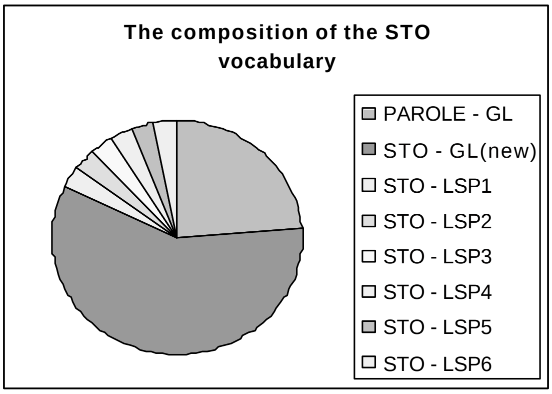

Figure 1: The composition of the STO vocabulary

5.3. Linguistic description – point of departure

The linguistic information content of the STO lexicon is organised according to the traditional practice in computational linguistics into three independent but coherently linked layers, i.e. the morphological, the syntactic and the semantic layer. Each descriptive layer is made up by a comprehensive system of the characteristic linguistic properties. Consequently, the full linguistic description of a lemma is structured in different sets of information, the so-called units. The representation model is based on a concept of such units. From the linguistic point of view a unit represents a particular linguistic behaviour of a lemma at the layer concerned, thus the full linguistic description of the lemma comprises morphological, syntactic and semantic units.

The STO structure shown differs from the PAROLE model in that it contains the lemma as underlying unit. This allows of linking on one hand more than one morphological units (e.g. spelling or inflectional variants) to a lemma, on the other hand the syntactic units are not linked to the morphological unit(s) but to the lemma itself. This solution provides an easy access to the independent layers. From the computational point of view a unit is a structured object containing a feature-based description expressed in attribute/value pairs. The linguistic information is divided up into fine pieces, i.e. many combining features. This approach ensures both flexibility and consistency in the linguistic description.

An efficient instrument to describe predictable and systematic behaviours is the use of patterns. Each pattern is a unique combination of several attribute/value pairs. The set of combining features is dependent on the layer of description and part-of-speech of the lemma. At the morphological level each pattern represents one particular inflectional behaviour. At the syntactic level a pattern describes a particular syntactic structure compatible with the lemma comprising features related to subcategorisation properties and raising/control phenomena.

In the process of extending the linguistic coverage we performed various development tasks pertinent to the degree of detail in inflectional and syntactic patterns.

5.4. Extending the linguistic coverage

In this section we focus on a few selected topics having particular relevance for the current work on the Danish lexicon. The linguistic coverage of the lexicon is extended in two ways. Firstly, further information types and features are added in order to implement more language-specific information. Secondly, the extensive use of corpus evidence provides increases the quantity and quality of the material to be coded.

5.4.1. Morphology – current developments

In STO, the general architecture and methods of the PAROLE descriptive model were adopted as a point of departure. This means that the concept of morphological units, their properties and the basic methodology of describing inflectional morphology are kept compatible with the principles common to all PAROLE lexicons. The current developments of our project are mainly concerned with the refinement and extension of the language-specific descriptions. In what follows we will discuss the developments concerning the morphology of nouns.

Selected tasks of current development of the morphological level:

?? Elaboration of new patterns of less frequent paradigms

?? Development of a systematic strategy for the treatment of words having inflectional alternatives

?? Addition of new features in order to treat language specific phenomena, such as
- compounding
- particular agreement cases

5.4.2. New patterns

The inflectional behavior of lemmas is described in patterns that are based on the 'remove/add' computational method, which is used to calculate inflectional forms of a l lemma. For nouns, the four basic forms are: singular indefinite, singular definite, plural indefinite and plural definite The definite forms are generated by adding the end-form article a suffix (see e.g. Allan et al. 1995; Underwood et al.1996/2001)

Briefly formulated, an inflectional form is calculated in two steps. (1) Remove the part of the lemma string, which does not remain unchanged when the particular inflected form is generated: this leaves the radical pertinent for the form. (2) Add the ending which generates the particular inflected form (which is not necessarily only a suffix in traditional sense) to this radical.

Example 1: *tale +n,+r,+rne*
The lemma is *tale* (sing. indef.; 'speech'); there is nothing to remove; the rule generates the following forms *talen* (sing. def. common), *taler* (plur. indef.) and *talerne* (plur.def.) by adding the appropriate endings.

The rule looks a bit more complicated when a part of the lemma has to be removed (in square brackets) for two of the inflected forms.

Example 2: *datter +en, [atter]øtre, [atter]øtrene*
This pattern generates from the lemma *datter* ('daughter') the following forms: *datteren* (sing. def. common); *døtre* (plur. indef.) and *døtrene* (plur. def.).

The information covered by a pattern includes both general types, such as number and gender and language specific types e.g. end-form definiteness of nouns, vowel dropping (syncope) and doubling of the final consonant in inflected forms. Each pattern is a unique combination of attributes and values, and a lemma may be linked to more than one single inflectional pattern.

During the newly finished encoding phase we detected a number of attribute/value combinations not yet covered. The appropriate new patterns (approx. 60) are now established and we regard the system almost fully developed for nouns. The STO lexical database contains presently 280 patterns of noun declension. (The figures for the other word classes are: pronouns 49, adjectives 75, verbs 155 and 'fuzzy' grammatical categories 10 patterns.)

5.4.3. Systematic treatment of inflectional alternatives

The development of new patterns is time-consuming and the 'cost/benefit' ratio less positive if only a very few lemmas belong to the pattern concerned.

This is often the case when encoding entry words of foreign origin (loan words), especially of Greek/Latin. A further difficulty is presented by their inflectional alternatives: several Danish variant forms, Greek and Latin forms are alternating. The Official Danish (Retskrivningsordbogen 2001), i.e. latest edition in electronic version – henceforth RO2001, does not provide any effective descriptive system for this task that could be exploited automatically. Danish grammars describe these forms as irregularity in the inflectional system (e.g. Allen 1995).

The following example from RO2001 illustrates the very compressed representation of alternating inflected forms:
Example 3:
virus *sb.*, -sen *el.* -set, -ser *el.* virus *el.* vira, *bf. pl.* virusse(r)ne *el.* viraene.

The article contains the lemma (in bold face) and meta-language information (in italics) - about part-of – speech *(sb)*, alternation between forms marked with *el.* ('or') and a particular marking of the definite plural *bf.pl.* There are recorded 8 variant forms. The first two forms are singular definite forms (doubling of the last consonant, gender: common or neuter), the forms number three, four and five are plural indefinite forms (one Danish and two of foreign origin. The last three are plural definite forms (two Danish and one foreign form). A further problem is that until year 2000 forms without consonant doubling (both genders) also were correct. Thus, those forms occur frequently in contemporary texts too.

The strategy adopted for the treatment of inflectional alternatives of a lemma is based on the requirements of automatic recognition and generation applications. This means that all alternatives must be equally recognised but in generation some alternatives may be preferred to others. Consequently, we sort out the alternatives in distinct paradigms on the basis of the combining values of the features covered by patterns, hereby separating Danish alternatives (i.e. variations of common/neuter gender, with/without doubling of the last consonant, etc.), Greek and Latin forms. The next step is to establish patterns for all variant inflections. Finally, the lemma is linked to all relevant patterns thus the representation in STO is consistent with other lemmas having more than one inflection possibility.

In this particular case we end up with 6 patterns according to RO2001. If we want to cover all - until recently accepted - forms occurring in the corpus for recognition purposes, we have to associate the lemma *'virus'* with 10 inflectional patterns.

As regards this particular lemma, we investigated the frequency of forms. We observed an interesting tendency showing that in texts about information technology the Danish inflectional forms are prevalent, while in texts from the health domain the Latin/Greek inspired forms are preferred. However, many texts of general language mix all forms – more or less randomly, although the use of a combination of Danish alternatives is striking.

5.4.4. Treatment of compound formation

In Danish the most productive method of word formation process, new words are coined by combining two or more independently existing words into a new one.

The method of compounding in Danish is very similar to German (although the compounds trend not to be as long as in German). Many Danish compounds are the equivalent of English noun phrases of N+N type (e.g. *armbåndsur* 'wrist watch') or a noun phrase containing an 'of'-genitive (e.g. *formuefordeling* 'distribution of wealth').

It is of course impossible to list each existing and potential compound of a language, on the other hand for NLP applications it is relevant to treat compounds properly. In order to treat this demand in the STO lexicon, two new features are introduced which extend dynamically the linguistic coverage of the lexicon.

First, compounds that are frequent in the corpus are inserted as entry words into the lexicon. Each compound entry has a separate field containing its decomposition into its immediate constituents, which are independent lemmas and linking element(s). The linking of constituents into compounds cannot be described sufficiently by applying general rules.

Example 4: *mandegruppe* => mand + e + gruppe (men's group)

Example 5: *børnehjem* => b[ørne]arn + hjem (children's home)

Remove/add computational rules are used to restore the lemma form of the first constituent (cf. section 5.4.2.)

Second, the prototypical linking element(s) are registered for all simplex nouns and lexicalised compounds as both lemma types may potentially appear as the first immediate constituent of compounds. The linking elements are ordered according to their frequency.

Example 6*: mand +e, +s,* +0 ('man')

Different linking elements are realised in compounds such as *mandemåned* ('man-month'), *mandsperson* ('male individual') *manddag* ('man-day').

Also the individual linking properties are registered consistently with the 'remove/add' computing method. We expect that these language specific extensions to the general linguistic descriptions will contribute considerably to a dynamic exploitation of the lexical resource.

5.4.5. Treatment of particular agreement cases

In automatic generation of texts, such as machine translation and abstracting one of the central tasks is to treat morphosyntactic agreement properly. In STO we observed that the rules applying to appellatives are insufficient to cope with geographical and geopolitical proper nouns, such as names of cities, mountains, rivers, seas and countries.

On the one hand, proper nouns semantically refer to a designated entity, which differentiates them from appellatives (common nouns). On the other hand, they constitute a subcategory of nouns having the basic morphological properties of nouns: gender, number, case and definiteness. However, it is not a trivial task to record the combination and appearance or absence of these features. Nevertheless, the morphosyntactic agreement is also required for these proper nouns along the same lines as for appellatives and it cannot be described by general language rules. This fact calls for a particular type of inflectional patterns describing geographical and geopolitical proper nouns. A pattern of this type records explicitly the characteristic inflectional restrictions e.g. form of definiteness (end-form article fixed or not), difference between the logical (semantic) and formal (morphological) number which affects the noun/adjective agreement conditions.

The STO lexicon contains 620 frequently used geographical and geopolitical proper nouns covering all different attribute/value combinations. The number of patterns elaborated specially for this purpose is 15.

5.5. Syntax

Recently we started working on some particular topics in Danish that require special attention at the syntactic layer. The development process is subdivided into two main steps. Firstly, we focus on a refinement of syntactic patterns and reconsider the appropriate degree of details of linguistic information to be given from the application point of view. Secondly, we revise the strategy for selection of patterns and description of syntactic units of a lemma. To this end the corpus-driven approach and relevant lexicometric principles are adopted.

The most substantial questions regard the selection of patterns to be represented for a lemma and 'birth level' of a unit. Both questions involve syntactic and semantic aspects as well. In this sense, the strictly modular representation model has some drawbacks because it is not always feasible to separate morphological, syntactic and semantic phenomena. Phrasal verbs, reflexive constructions and collocations are complex morphosyntactic structures but units at the semantic layer having one single lexicalised meaning. The principles behind the design of patterns for collocations are described in Braasch & Olsen 2000. In order to prepare these structures for a proper linking to units of the semantic layer, the syntactic descriptions will be supplied with a few features treating the internal structure of these complex items.

6. Further developments

For the time being the STO project was mainly concerned with morphological tasks and also with a number of implementations for the syntactic layer, as described above. Figure 2 (below) shows the present state of the STO lexical database.

At the semantic layer the information will be provided at three specificity levels. The vocabulary of specialised languages will only be coded with domain information (Level 1). Most of the general language vocabulary will be described with more comprehensive information, namely sense distinctions, ontological typing and selectional restrictions (Level 2). This means that level 1 and 2 represent a relatively lean semantics.

A further, extensive and rich Level 3 of specificity is implemented in the SIMPLE project (Pedersen & Keson 1999) which covers the semantic description of approx.10,000 entry words of PAROLE/STO. This level contains detailed information about semantic relations, argument structure, information, such as ontological typing, domain information, qualia structure, semantic relations, event structure'. At a longer term it is planned to supply a larger part of the STO lexicon with level 3 semantics – however, this extension is depending on an additional funding of the project.

7. Exploitation of the STO lexical resource

The material in the lexicon consists of a large number of small information pieces stored in well-structured database tables. The production of entries is subdivided into production modules. Each module is uploaded into the central database after validation.

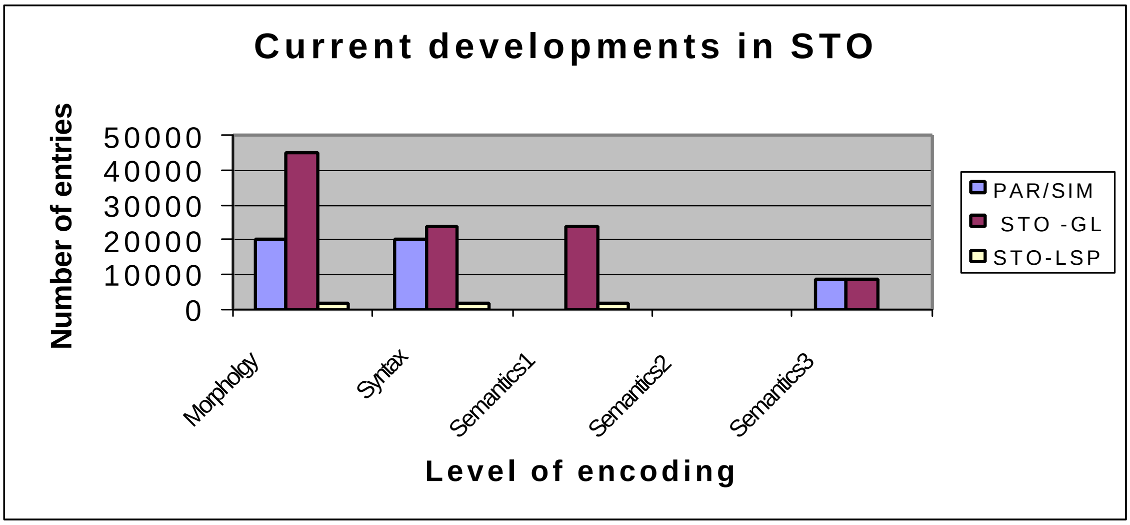

Figure 2: Overview of current STO

At present, according modules of the lexicon material are ready to be downloaded in a general format. The process is to be carried out by CST's database administrator on demand. The material can then be customised for a particular application e.g. it can be extended with some additional information required by a particular application or processing software. These processes can be performed on the lines laid down in the Danish national standard for lexical data collections (STANLEX 1998) concerning the classification of information types and draft for information structures.

For the longer term, we intend to develop appropriate tools in order to make this process easier and speed it up. The customers will be able to define their own needs and requirements in a predefined structure reflecting the structure and information content of the STO database. The customer's specification will be used as input for a selection and downloading-on-demand procedure carried out automatically.

Presently, we are working on the development of an internet-based user-interface, which will provide a browser access to the lexical database. A demonstration module will be made available later this year on the web address www.cst.dk/projects/sto. Guidelines for the user, underlying documentation of the database and linguistic specifications will also be electronically accessible.

8. Summing up

The current developments of the Danish lexicon showing that the STO project successfully exploits the experience acquired and the lexical resource produced within the framework of the LE-PAROLE project. The main work performed so far concentrated on language-specific refinements and extension of the lexical coverage.

Although the STO project focuses on the monolingually relevant information content and data structure, we are also aware of the need for a Danish lexicon that can be integrated into multi-lingual lexical resources. To this end, the lexical data produced are kept compatible with the PAROLE descriptive language and as regards the semantic layer we remain attentive to structures produced within other follow-up projects, like SIMPLE.

The work presented here suggests a number of additional tasks, for example we plan to explore more advanced statistical methods in evaluating corpus evidence in order to improve the lexical and linguistic coverage of the STO lexicon.

9. Acknowledgements

Many thanks go to the members of the STO project group: Bolette S. Pedersen, Sussi Olsen, Sanni Nimb, Nicolai Hartvig Sørensen, Dorthe Haltrup Hansen and Anna Sophie Liebst for their valuable contributions to this article. Furthermore, the STO project gratefully acknowledges the support given by colleagues at The Danish Dictionary (DDO) project.

10. References

Allan, R., P. Holmes, T. Lundskær-Nielsen (1995). Danish: A Comprehensive Grammar. Routledge, London/New York.

Bergenholtz H. & S. Tarp, eds. (1995). Manual of Specialised Lexicography. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Braasch, A., C. Navarretta & N. H. Sørensen (1998). LE-PAROLE. Danish Lexicon Documentation. (Project document) 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--10017). Athens, Greece

Guimier, E., A. Ogonowski & PAROLE Partners (1998). LE-PAROLE. Report on the Morphological Layer. (Public project document) Paris.

Guimier, E., A. Ogonowski & PAROLE Partners (1998). LE-PAROLE. Report on the Syntactic Layer. (Public project document) Paris.

Lebart L., A. Salem and L. Berry (1998) Exploring Textual Data. Kluwer Academic Publishers, Dordrecht/Boston/London.

Navarretta, C. (1997): Encoding Danish Verbs in the PAROLE Model. In *Proceedings from RANLP '97*. Tzigov Chark, Bulgaria

Olsen, S. (2002). Lemma selection in domain specific computational lexica – Some specific problems. Forthcoming - In *Third International Conference on Language Resources and Evaluation*. Las Palmas, Spain

Pedersen, B.S. & B. Keson (1999): 'SIMPLE - Semantic Information for Multifunctional Plurilingual Lexicons:

Some Examples of Danish. Concrete Nouns. In *SIGLEX 1999*, *ACL-Workshop*, Maryland, USA.

STANLEX (1998). DS 2941-1. Leksikalske datasamlinger. Indholds- og strukturbeskrivelse. Del 1. Taksonomi til klassifikation af oplysningstyper. Dansk Standard. København.

Underwood, N., C. Povlsen, P. Paggio, A. Neville, B. Pedersen, B. Ørsnæs & A. Braasch (1996/2001). LINDA, Linguistic Specifications for Danish

Electronic versions of The Official Danish Spelling Dictionary:

Retskrivningsordbogen (1986). Dansk Sprognævn., Copenhagen

Retskrivningsordbogen (2001). Dansk Sprognævn, Copenhagen