Danish Monolingual Lexicon

Documentation
Version 2

Author:	Anna Braasch, Costanza Navarretta, Sussi Olsen, Bolette S. Pedersen
Editor:	Anna Braasch
Institute:	Center for Sprogteknologi (CST), University of Copenhagen
Address:	Njalsgade 80, DK-2300 Copenhagen S, Denmark
Email:	anna@cst.dk sussi@cst.dk
Date:	25/02/2008
Version:	2

PART 1:

GENERAL DESCRIPTION OF THE LEXICON

DESCRIPTION OF THE MORPHOLOGICAL LAYER

1 General introduction

The STO (SprogTeknologisk Ordbase) lexicon is a comprehensive computational lexicon of Danish developed for NLP/HLT applications. STO is created within the framework of a national collaborational project, initiated by Center for Sprogteknologi (CST). The work was founded on a contract with the Danish Ministry for Science, Technology and Development. The duration of the project was three years, ending by February 2004.

The lexicon material is produced by the following project partners:
- Center for Sprogteknologi, University of Copenhagen,
- Institut for Datalingvistik, Copenhagen Business School,
- Institut for Almen og Anvendt Sprogvidenskab, University of Copenhagen
- Institut for Fagsprog, Kommunikation og Informationsvidenskab, University of Southern Denmark.

Contact persons:
Hanne Fersøe, Deputy Manager e-mail: hanne@cst.dk (marketing)
Anna Braasch, Senior Researcher e-mail: anna@cst.dk (database contents)
Costanza Navarretta, Senior Researcher e-mail: costanza@cst.dk (XML support)

About this documentation

The documentation consists of two parts:
Part 1: General description of the STO lexicon and Documentation of the Morphological Layer
Part 2: Documentation of the Syntactic Layer

The present Part 1 contains all relevant general and background information and the description of the morphological layer.
All information about the Syntactic layer is provided in Part 2.

The list of references provided contains not only the literature referenced within this documentation, but also a few publications which may be relevant for the user.

2 Technical specifications

2.1 Description of the data files extracted from the STO database

The entire lexicon comprises two layers of description: the morphological layer where the units are provided with morphological description (A), and a the syntactic layer where the units are provided with syntactic information (B).

The deliverable of lexicon data is split into two standard packages
- Deliverable A: the Morphological Layer
- Deliverable B: the Syntactic Layer

Accordingly, the documentation is split into two parts, as mentioned above.

2.2 Deliverable A: Morphology

The morphological layer of the lexicon contains a vocabulary of 81,524 entry words with comprehensive morphological descriptions. The selection of entry words and the description method is documented in Chapter 3.

The morphological lexicon is per default provided in a comma-separated values (CSV) file format, which allows for import of data into various formats, e.g. into a mysql table.
The morphological lexicon is subdivided into 10 part of speech files and one word form file with frequency information. The directory with the data files contains a README-file with file names and file sizes.
The specifications for the ten part of speech files and the frequency file are enclosed in the appendices of this document.

Number of Files	**Content**	**No. of entries**	**File size in bytes**	**Specification file in appendix**
1	Nouns	64,735	121311721	Appendix A
1	Verbs	9,773	1147652	Appendix B
1	Adjectives	5,775	1505287	Appendix C
1	Adverbs, Prepositions, Conjunctions, Interjections, Unique	1,197	54376	Appendix D
6	Pronouns - demonstrative - indefinite - interrogative - personal - possessive - reciprocal	44 in total	 327 751 326 523 709 138	Appendix E
1	Word forms with frequency	692410	52653347	Appendix F

2.3 Deliverable B: Syntax

The syntactic layer contains detailed syntactic description of 45,000 entry words of the vocabulary mentioned above.

The syntactic lexicon is provided in the extended mark-up language (XML) file format and the material is subdivided into a number of files in order to deliver manageable file sizes.

The data material is provided as three XML files as follows (size in bytes):

STO_Syntax_1_v1.xml	4437723
STO_Syntax_2_v1.xml	4872856
STO_Syntax_3_v1.xml	4488728

The data files can be validated with the XML Schema which can be found in Appendix 1. (File name: STO_Syntax.xsd, size 21865 bytes).

For a detailed description of the syntactic lexicon see Part 2, Documentation of the Syntactic Layer.

3 Lexicon Description

3.1 Background

The establishment of the descriptive model and the linguistic specifications for STO greatly benefits from the experience acquired at CST within the framework of the multi-lingual LE2-4017 - 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. This project was aimed to the development of re-usable language data, i.e. corpora and electronic lexica in all languages of the European Union. The goal of the project was to produce for the languages involved (1) a corpus of 20 million running words and (2) a lexicon of 20.000 entries. The Danish PAROLE lexicon was produced by CST.
The PAROLE lexicons were built around a generic model (an instantiation of the EAGLES recommendations in an enriched GENELEX model). (For further information please consult the Executive summary of the LE-PAROLE project: www.hltcentral.org/usr_docs/ project-source/parole/ParoleFinal.pdf).

3.2 Contents of the lexicon

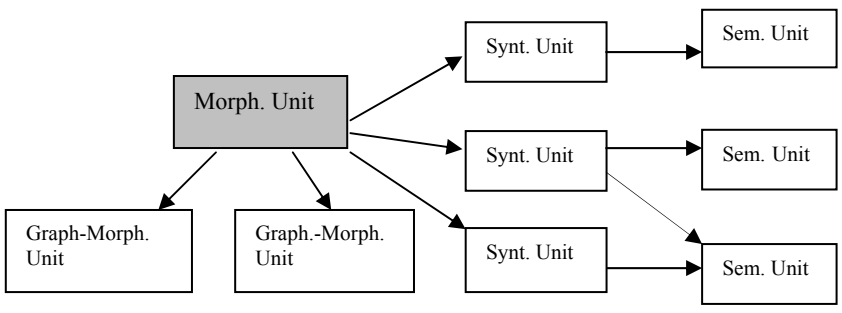
3.2.1 Linguistic description: Method and model

The STO lexicon is corpus based both as regards 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 into three independent descriptive 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. 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 full linguistic description of a lemma comprises a set of morphological, syntactic and semantic units. These units are, although independent, encoded in a coherent way, and they are linked together in the central STO database providing the linguistic description of a lemma. The representation model underlying the STO lexicon is based on a concept of units and the links between them.

The STO model of description



3.3 Composition of the lexicon

The STO lexicon contains over 81,000 lemmas, of which approx. 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 also with a fine-grained syntactic description as well. The tables (1 through 3) below show the composition of the vocabulary covered in detail. 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. We consider this to be a kind of transitional area between the general language and specialised expert languages.

3.4 The coverage of the lexicon

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 it shows also to which extent the different word classes have been provided with a) only morphological information, b) with morphological and syntactic information.

3.5 General language and domain language vocabulary

Lexical Category	**No. of Lemmas**	**Morph. only**	**Morph. & Synt.**
Noun	**64735**	47%	**41%**
Adjective	**9773**	32%	**55%**
Verb	**5775**	2%	**81%**
Adverb	**771**	81%	**19%**
Interjection	**158**	100%	**0%**
Preposition	**80**	100%	**0%**
Conjunction	**60**	100%	**0%**
Pronoun	**44**	100%	**0%**
Misc.	**128**	100%	**0%**
Total	**81524**		

Table 1: The vocabulary of the STO lexicon in total

Table 2 contains the figures for the general language vocabulary, all closed word classes belong to this category.

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 in the STO database with part of speech distribution

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

3.5.1 Representation of closed word classes

The following closed word classes (function words) are covered exhaustively, viz. registered by their lexical category at the morphological layer:

- Pronouns: subclasses: personal, possessive, relative, demonstrative, interrogative, indefinite Adpositions: Prepositions (which make up the only subclass in Danish)
- Auxiliary verbs
- Conjunctions
- Infinitive marker
- Unique
- Interjections (registered to a large extent but possibly not fully exhaustively because of the fact that this class is slightly productive).

3.6 Description of the General language and domain language corpora

3.6.1 General language: corpora and lemma selection

The lemma selection and the linguistic description of the entire STO vocabulary are mainly based on text corpora composed for other purposes. As regards the general language coverage, the selection of lemmas takes as its starting point a frequency based provisional lemma list of approx. 200,000 lemma candidates. This list was originally compiled for The Danish Dictionary (DDO) by the Danish Society for Language and Literature. A corpus of modern Danish (time period: 1983 – 92, size approx. 36 M tokens) served as a basis for this provisional list. Subsequently, it has been manually revised for STO and supplemented on the basis of other corpus resources, viz. a newspaper corpus (Berlingske Tidende, year 1999). This final list contained approx. 68,000 general language words, selected by frequency. Since 2002, two corpora, the Korpus 2000 and Korpus 90 are on-line and freely accessible at http://korpus.dsl.dk/korpus2000 . Thus, in the last phase of the project also these corpora were consulted for control and referencing purposes.

Overview of the general language corpora

Corpus	Size	Composition	Topic examples	SELECTION
Berlingske Tidende &	30 M	Newspaper articles and	Domestic and foreign affairs, economics, administration, law,	A full volume of the daily and

Weekendavisen (1999)		reports in full length	sport, culture, consumption, amusement, gardening, etc.	weekly newspaper exclusive the advertisement sections
DK87-90 (time period: 1987-89)	4 M	Newspapers, periodicals, magazines, books,	Fiction, popular science, everyday life …	Text samples of limited size; the text selection is based on a principled corpus design
Korpus90 (time period: 1988-92)	28 M	Part of the DDO corpus; Books, magazines, newspapers	A broad range of general topics as described in daily newspapers, periodicals, magazines, fiction, personal letters, transcribed conversations and speeches	Text samples of various length; the text selection is based on a thorough corpus design
Korpus2000 (time period: 1998-2002)	28 M	Around the Year 2000: Books, magazines, newspapers	A broad range of general topics as described in daily newspapers, periodicals, magazines, fiction, personal letters, transcribed conversations and speeches	Text samples of various length; the text selection is based on a thorough corpus design

Table 4. General language corpora, size and text types

Selection of general language lemmas

Initially, a lemma candidate list has been set up on the basis of a lemma list from the Danish Dictionary (DDO) project, whereof lemmas having a frequency above 20 have been selected for STO. In the second run, the list of candidate lemmas has been verified by searches in a newspaper corpus. Further, general language words occurring in the domain texts selected (cf. below) have been added to the lemma list.

3.6.2 Domain languages: corpora and lemma selection

In order to enlarge the coverage of the lexicon also lemmas from domain language texts are included.

The domain-related vocabulary has been selected from six domain specific corpora each of them having a size between 1 and 2 M million tokens (cf. below, Table 5). These corpora are collected from various on-line resources, mainly from public information websites and the texts selected are mainly originating from communication written by experts to laymen. The lemmas extracted were not highly specialized terms but rather words that belong to the everyday communication about a particular domain thus being in the grey area between general and domain expert languages.

Method of text collection

The method and the process of collecting texts for the linguistic investigations and the editing of the lemma candidate lists were to a high degree automatic. The text selection was based on the so-called onomasiological approach, which means that the definition and delimitation of the domain was based on central topics of the domain in question. "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 the domain." (Jørgensen et al., 2003). The OS served as a basis for establishing the collection of web documents. The items from the OS were then used as search words to identify relevant texts on the web covering at least one, or preferably more, topics of the domain. This approach was intended to guide the selection of the corpus with a sufficient *coverage* of the domain, but without weighting. The method is used to good advantage in reducing the risk of circularity between search words selected and texts identified. For a further discussion of the building of domain specific corpora cf. Jørgensen (op.cit.)

These text collections also form the basis for the description of linguistic features. On the other hand, they only serve as a basis for investigations of language usage below the sentence level. Thus, the texts cannot be reconstructed or exploited for other purposes.

Overview of the domain text collections

Domain (Danish Corpus name)	**No. of Tokens**	**Examples of Text types**	**Examples of Topics**
IT (EDB-KORPUS)	1.1 M	Technical and popular magazine articles; textbooks	Hardware, software, CPU, external devices, operating system, programming language,
Environment (MILJØ-KORPUS)	1.5 M	Public information from Ministry of the Environment, relevant authorities, organizations (Greenpeace)	Environment control and policy, environmental planning and management, energy, working environment, exposure, pollution of waters, earth and air
Commerce (H&E-KORPUS)	1.5 M	Public information from the Ministry of Finance, Public services, relevant authorities and organisations	Distribution, foreign trade, commerce, business management, export, import, sales, marketing, legislation for commerce, restrictions on trade
Public Administration (FORVALT-KORPUS)	2.6 M	Public information from the Government services and authorities, organizations	State, county and municipality administration, public institutions, public employees, public administration, taxation
Health (SUNDHEDSKORPUS)	1.1 M	Public information from health department and sanitary authorities; medical records, case reports, answers to FAQs	Health services, hospital service, nursery, nutrition, preventive and alternative medicine, patient treatment, health insurance
Finance (FINANSKORPUS)	1.9 M	Public information from authorities, organizations; short on-line instructive and informative publications	Economics, macro - & micro economy, financial structures, markets, tasks, laws and organisations

TOTAL	**9.7 M**		

Table 5. Collections of domain texts (corpus)

The IT texts originate from 1997 to 2000; all other domain text collections are compiled during the time period 2002 – 2003.

Selection of domain specific lemmas
A lemma candidate list was generated automatically after the tokenization and lemmatization of the corpus. This list was a result of a comparison between common language words already encoded in the STO database and the full lemma list of the domain corpus. We observed a drawback of this simple comparison method, namely words having both a general language reading and a domain specific reading are not picked for the lemma candidate list if they already were encoded, e.g. *mus* 'mouse', with a common and a computer-related reading (IT domain).
From the lemma candidate list were manually selected the relevant domain specific lemmas (with a frequency higher than 2), in this process also errors in the POS-tagging and lemmatization were corrected.

The following candidates were not selected for STO:
- Proper names
- Expert terms
- Long and unusual compounds
- Misspellings and other errors (e.g. candidates being overrepresented owing to identical documents in the corpus)

General language words appearing on the candidate list are encoded as such.

The table below summarizes the main steps of the lemma selection.

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

Table 6. Domain specific lemma selection (Source: Jørgensen op.cit.)

3.7 The alphabet of Danish

The alphabet of Danish comprises 29 legal characters; each of them is in principle to be found in every position within words. However a few of them appear only in words of foreign character (viz. *q, w, z*.) Each of the characters can appear both in lower and in upper case.

The characters in alphabetic order are:

a b c d e f g h i j k l m n o p q r s t u v w x y z æ ø å
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z Æ Ø Å.

Notes
For the characters *æ*, *Æ*, *ø, Ø,* and *å, Å,* there exist obsolete spelling alternatives, viz. *ae, Ae, oe, Oe,* and *aa, Aa*, resp. These variants are not included in STO, although they are legally used in family names e.g. *Bjerregaard, Kjaergaard, Selsoe* and in a few other cases, e.g. brand names based on geographic names such as *Aalborg Akvavit*.

In some texts written in foreign languages containing Danish words, these spelling alternatives are still used on occasion not only in names but in other words too, because of the fact that keyboards don't have these characters as a standard.

The string CO[2], read CO subscript2

4 The linguistic content of the lexicon

The linguistic description of a lemma is subdivided into three layers, viz. the morphological, syntactic and semantic layer. According to this approach, the entire lexicon consists of description units of these levels: morphological, syntactic and semantic units. In the following, we describe the linguistic information represented at the respective layers. The structure allows for linking on one hand more than one graphical units (viz. spelling or inflectional variants) to a single morphological unit, on the other hand the syntactic units are not linked to the graphical unit(s) but to the morphological unit 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.

4.1 Orthography

4.1.1 Spelling and variants in STO

There exists for Danish an Official Spelling Dictionary (Retskrivningsordbogen, henceforth abbreviated RO). The current version is updated in 2001 (henceforth RO2001). The present material contains not only forms that are in accordance with RO2001 but also some obsolete spelling variants and inflectional forms originating from the period between 1986 and 2001. The reason for including these variant forms in the lexicon is the fact that they are useful in recognition processes. The feature RO-approved with the values 'yes', 'no' is employed to mark the validity of spellings, spelling paradigms and specific inflected forms, which makes it possible to prevent their use in generation processes. The latest update of the STO material is in accordance with the latest spelling norm RO 2001.

4.1.2 Spelling and inflection of new words with foreign origin

When encoding entry words of foreign origin (loan words), we met spelling variants and inflected forms in the corpus, which are not (yet) registered in RO2001. All these forms have been approved through consultation with the Danish Language Council. Also words originating from domain texts presented some difficulties because of a number of inflectional alternatives, gender selection and syntactic construction as well. 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.

4.2 The morphological layer

The table below shows the distribution of entry words in the lexicon among the various categories/subcategories. Very few words are not encoded with lexical category, (WithoutC = without lexical category) and a few categories are not subdivided into subcategories (WithoutSC= without subcategory.)

Lexical Category	Lexical Subcategory	Morphological Units	Example
NOUN	COMMON	64131	abonnent
NOUN	PROPER	604	Abessinien
VERB	MAIN	5719	adressere
VERB	MEDIAL	56	lykkes
ADJECTIVE	NORMAL	9651	god
ADJECTIVE	CARDINAL	72	atten
ADJECTIVE	ORDINAL	50	attende
PRONOUN	DEMONSTRATIVE	5	begge, den
PRONOUN	POSSESSIVE	11	din, dens
PRONOUN	RECIPROCAL	2	hinanden
PRONOUN	INTERROGATIVE	5	hvad
PRONOUN	PERSONAL	10	de, sig
PRONOUN	INDEFINITE	10	alting, en
ADVERB	GENERAL		ofte
ADPOSITION	PREPOSITION	80	uden for, på
CONJUNCTION	WITHOUTSC	60	bare
INTERJECTION	WITHOUTSC	158	adjø
UNIQUE	WITHOUTSC	1	som
UNIQUE	FORMALSUBJECT	1	der
UNIQUE	INFMARK	1	at
WITHOUTC	WITHOUTSC	125	a conto

Table 7: Lexical categories in STO

The basic unit of this layer is the *Morphological Unit (MU),* which identifies the entry word providing a unique identifier (Mu_id), lexical category and a few other, mainly administrative information types. Thus, the morphological unit functions in most respects like a lemma or entry word in editorial dictionaries, i.e. the whole set of information can be accessed by the morphological unit. Of course, a database structure allows for several other access paths.
The main unit of morphological description is the *Graphical Morphological Unit (GMU)*, which is provided with information on spelling, inflection, compounding/decomposition. A morphological unit can have more than one spelling variant or inflectional variant, thus it can be linked to more than one single GMU.

This layer concentrates on the following general information types
1. Linguistic information types

- Lexical category (part of speech)
- Spelling (the basic form of the entry word)
- Inflection (if applicable)

2. Other information types

- Approval (of orthography, cf. below)
- Origin (i.e. the source from where the entry word has been selected; general language words can have two different sources, domain words originate from the various domain corpuses)
- Frequency based on the two mayor Danish corpora of general language, Korpus90 and Korpus2000.

In addition, there may appear some linguistic information, which is specific to a particular category or subcategory such as word formation, viz. compounding (only for nouns) or transcategorization (for adjectives and verbs), and inflectional agreement features for geo-political proper nouns.

4.2.1 Treatment of homographs

Homograph lemmas having identical lexical category, graphical inflectional paradigm (GINP) and joining element (cf. below, 'Fugeelement') are encoded as one single morphological unit because there is no morphological difference observed between them, although they have different meanings.

Ex.: *pande* (noun) 'pan'; 'forehead',
Encoding: MU_ID: *pande_1*; inflectional pattern for both: GINP_ID: MFG0076 (+n,+r,+rne)

Homograph lemmas showing morphological differences in their lexical category, inflectional paradigm and/or joining element) are encoded as distinct morphological units.

Ex: (a) *skade*, noun, ('skate'/ 'magpie'; or 'damage'/ 'injury')
(b) *skade*, verb, ('damage'/ 'injure')
Encodings for (a):
MU_ID: *skade_1*; inflectional pattern GINP_ID: MFG0076 (+n,+r,+rne) (for 'skate')
Joining element ('Fuge'): Removed: Added: 0 Result: "skade"
MU_ID: *skade_2*; inflectional pattern GINP_ID: MFG0076 (+n,+r,+rne) (for 'damage')
Joining element ('Fuge'): Removed: Added: 0 Result: "skade"
Joining element ('Fuge'): Removed: Added: s Result: "skades"

Encodings for (b):
MU_ID: *skade_3*; inflectional pattern GINP_ID: MFG0112 (V:INF:+,+s,PRE:+r,+s,P…)

4.2.2 Treatment of spelling variants

A rather limited number of lemmas have more than one single spelling; these are encoded as alternative spellings of the morphological unit in question, as follows:

Ex: *hæfte* or *hefte* 'booklet'
Encoding: MU_ID: *hæfte_1*
Gmu_id*: GMU_HÆFTE,1_1*
Spelling*: hefte*
Gmu_id*: GMU_HÆFTE,1_2*
Spelling*: hæfte*

Some alternative spellings are frequent spellings that are not approved by the Danish Language Council in RO 2001. These appear with a 'NO' for RO_Approved.

4.3 Morphological information

This section describes the features encoded in the following way: For each category (in other terms part of speech or word class) we list the relevant linguistic features and their respective lists of legal values. Relevant language specific notes and illustrative examples are given after the entire list.

NOUN

- **Subcategory**: *common, proper.*
- **Gender**: *common, neuter, unmarked*.
- **Number**: *singular, plural.*
- **Case**: *genitive, unmarked.*
- **Definiteness**: *definite, indefinite, unmarked.*
- **Fugeelement** (joining element): *s, e, 0.*
- **Decomposition**: a string in the format*: noun + [insertion rule of fuge] + noun* or *noun + noun*

ADJ ECTIVE

- **Subcategory:** *normal, ordinal, cardinal.*
- **Number**: *singular, plural.*
- **Gender**: *common, neuter.*
- **Definiteness:** *indefinite, definite.*
- **Function:** *attributive, predicative.*
- **Degree**: *positive, comparative, superlative*.
- **Transcat**: *transadverbial*

VERB

- **Subcategory**: *main, medial.*
- **Mood**: *infinitive, indicative, imperative, gerund, participle.*
- **Tense**: *present, past.*
- **Voice**: *active, passive.*
- **Transcat:** *transnominal, transadjectival*

PRONOUN

- **Subcategory:** *personal, demonstrative, indefinite, interrogative, reciprocal, possessive.*
- **Number**: *singular, plural.*
- **Gender**: *common, neuter, unmarked.*
- **Person**: *1, 2, 3.*
- **Possessor:** *singular, plural*
- **Case:** *genitive, unmarked.*
- **Register**: *formal.*

ADVERB

- **Subcategory**: *general*

ADPOSITION

- **Subcategory:** *preposition*

UNIQUE

- **Subcategory:** *infinitive marker, formal subject.*

CONJUNCTION

INTERJECTION

4.4 Inflectional behaviour

The most basic morphological information type concerns the inflectional behavior dealing with the variation in form of words for grammatical purposes.

4.4.1 Method of description

The information to be covered 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. A unique combination of relevant attributes and values make up an inflectional pattern (GINP), and a morphological unit (here also called lemma) may be linked to more than one single inflectional pattern.

The inflectional behavior of lemmas is described by employing the 'remove/add' computational method, which is used to calculate the particular inflected forms of a lemma. Briefly formulated, an inflected form is calculated in two steps:
(1) *REM*: 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*: Add the ending which generates the particular inflected form (which is not necessarily only a suffix in traditional sense) to this radical.

Examples
For nouns, the four basic forms are: singular indefinite (the usual lemma form), 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) to the appropriate indefinite form.

Example 1: *tale +n,+r,+rne*
The lemma is *tale* (sing. indef.; 'speech'); GINP_ID: MFG0076 (in the example represented by its Naming which demonstrates the appropriate endings +n,+r,+rne) expresses the following generation rules: there is nothing to remove; the rule generates the following forms by adding the appropriate endings:

talen (sing. def. common)
taler (plur. indef.)
talerne (plur.def.).

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* GINP_ID: MFG0024 (+en,[atter]øtre,[atter]øtrene)
This pattern generates from the lemma *datter* ('daughter') the following forms:

datteren (sing. def. common)
døtre (plur. indef.)
døtrene (plur. def.).

The above forms are unmarked for case, all genitive forms are generated by a general rule by adding the suffix +*s* to the appropriate unmarked form.

4.5 Explanations and examples of word classes

The assignment of part of speech (word class) to the lemmas is in accordance with the Official Danish Spelling Dictionary (2001).

4.5.1 *Nouns*

Subcategories: Common nouns are appellatives (*bog* 'book'), the encoded proper nouns are mainly geo-political nouns (*Danmark* 'Denmark') and a few other types e.g. celestial bodies (*Venus*).

The morphological unit is identical with the primary (basic) form of a word, which is for nouns with full inflectional paradigm the singular, indefinite form, unmarked for case. Exceptions:
(a) For nouns lacking singular form (i.e. being pluralia tantum), the plural indefinite form is regarded as its primary form (*penge* 'money', *mæslinger* 'measles'). Though, a few of these nouns can appear in singular in particular texts of LSP (e.g. *bukser* 'trousers'). The gender of pluralia tantum nouns is *unmarked*.
(b) For nouns without indefinite form, the definite form is used (*Filippinerne* 'the Philippines').

The general description method is applied also to nouns without full inflectional paradigm as regards setting up an appropriate GINP, only the lacking forms are left empty.

The noun declension system in Danish is rather simple, only the genitive has an inflectional suffix, viz. *–s.* All other traditional cases (nominative, accusative, dative) are inflectionally *unmarked*.

Example: *dag* 'day', with full declension: for illustration purposes, the singular genitive suffix and end definiteness marker, and their combination are printed in bold face.

The table below shows the inflection features of a common noun having a full paradigm.

WORD FORM	GENDER	NUMBER	DEFINITENESS	CASE
dag	COMMON	SINGULAR	INDEFINITE	UNMARKED
dag**s**	COMMON	SINGULAR	INDEFINITE	**GENITIVE**
dag**en**	COMMON	SINGULAR	**DEFINITE**	UNMARKED
dage**ns**	COMMON	SINGULAR	**DEFINITE**	**GENITIVE**
dagene	COMMON	PLURAL	DEFINITE	UNMARKED
dagenes	COMMON	PLURAL	DEFINITE	GENITIVE
dage	COMMON	PLURAL	INDEFINITE	UNMARKED
dages	COMMON	PLURAL	INDEFINITE	GENITIVE

Table 8. Declension of a common noun

Fugeelement (joining element) information on both simplex nouns and shorter compounds
The joining element (*-s* or *–e*) follows the noun and is joined by another noun component to form a compound noun.
Ex:
Spelling: *ansvar* ('responsibility') Fugeelement: Removed: Added: ***s*** Resultat: *ansvars*
Compound noun: *ansvar**s**fordeling*

The Decomposition feature is only used for noun + noun compounds. It contains the segmentation of a compound noun into its two immediate noun components and the joining element in between them (if there is one), '+' is used as joint marker.

The *REM/ADD* method (described above, Method of description) is also applied for describing noun compound formation.

Example 3: *arbejdsdeling (*'division of labour', lit.: 'labourdivision')
Decomposition: arbejde+*[e]s*+deling

The format of the information given here can be
*noun + [*calculating rule for insertion of 'fugeelement'*] + noun*
noun + noun (if there is none).
Ex:
Spelling: *ansvarsbevidsthed*
Decomposition: *ansvar+**s**+bevidsthed* (lit: responsibilitysense, 'sense of responsibility').

4.5.1.1 Geo-political proper nouns
The morphological patterns of geo-political nouns cater also for their particular agreement features in order to facilitate proper generation.

Lemma	Definiteness suffix	Genus	Number	Article and attributive adjective	Predicative construction with adjective
Donau	-	com.	sing.	Den brede Donau	Donau er bred.
Tyskland	-	neu.	sing.	Det rige Tyskland	Tyskland er rigt.
København	-	neu.	sing.	Det store København	København er stor.
Rhinen	løs	com.	sing.	Den snavsede Rhin	Rhinen er bred.
Elben	fast	com.	sing.	Den brunlige Elben	Elben er bred.
Arresø	- (+en)	com.	sing.	Den varme Arresø	Arresøen er varm.
Sortehavet	fast	neu.	sing.	Det varme Sortehavet	Sortehavet er varmt.
Atlanterhavet	løs	neu.	sing.	Det kolde Atlanterhav	Atlanterhavet er koldt.
Atlasbjergene	løs	ø	plur.	De høje Atlasbjerge	Atlasbjergene er høje.
Filippinerne	fast	neu.	plur.	Det vestlige Filippinerne	Filippinerne er rigt på ressourcer.
Færøerne	fixed [region]	neu.	plur.	Det smukke Færøerne	Færøerne er rigt på vand.
	detachable [groupe]	ø	plur.	De 18 Færøer	Færøerne er smukke.
Christiansø	-	com.	sing.	Det smukke Christiansø	Christiansø er smuk(t).

Tabel 9: Overview of the agreement features of geo-political proper nouns (sample)

4.5.2 Adjectives

The lexical category of adjectives is subdivided into three subcategories: normal (*blid* 'gentle, kind, mild'), cardinal (*atten* 'eighteen') and ordinal (*attende* 'eighteenth'), cf. the Official Danish Spelling Dictionary, RO2001. The same work of reference is followed also in specific cases, where it from a functional point of view is difficult to assign the lemma unambiguously to a particular lexical category. The lemmas below have attributive and nominal use as well, which combine with different agreement features.
Thus,

- *al* is categorized as adjective, with subcategory normal.

The following are categorized as pronouns, with subcategory indefinite

- *ingen* (attributive function: ‘no, not any’; nominal function: ‘no one, nobody’) ,
- *enhver* (attributive function: ‘any, everybody’; nominal function: ‘anyone, everyone’)
- *nogen* (attributive function: ‘some, any’; nominal function: ‘somebody, someone’ and ‘something’, etc.)

The morphological unit of an adjective is identical with its basic form, viz. positive degree, common gender, singular, indefinite form (*blid*).

The adjective declension system comprises the following basic features: Adjectives are inflected in gender, number and definiteness.
Adjectives change form both in attributive and in predicative function, as required by the gender and number of noun/pronoun they describe. In the predicative function in singular, only the rule of gender agreement applies (i.e. there is no definiteness agreement). In plural, only the rule of number agreement applies in both functions (i.e. neither definiteness nor gender agreement). A few adjectives have only a basic form and are not inflected at all, e.g. *beige* ‘beige’.

The table below summarizes the basic agreement rules:

	Attributive function		**Predicative function**	
Agreement	Singular	Plural	Singular	Plural
Common, indefinite	En blid pige	Blide piger	En pige er blid	Piger er blide
Common, definite	Den blide pige	De blide piger	Pigen er blid	Pigerne er blide
Neuter indefinite	Et stort hus	Store huse	Et hus er stort	Huse er store
Neuter definite	Det store hus	De store huse	Huset er stort	Husene er store

Table 10. Adjective phrases, basic agreement rules

The table below shows the inflection features of an adjective (normal) with full paradigm. For illustration purposes, the suffixes are highlighted.

WORD FORM	GENDER	NUMBER	DEFINITENESS	TRANSCAT	FUNCTION	DEGREE
blid	COMMON	SINGULAR	INDEFINITE		ATTRIBUTIVE	POSITIVE
blid	COMMON	SINGULAR			PREDICATIVE	POSITIVE
blid**t**	NEUTER	SINGULAR	INDEFINITE		ATTRIBUTIVE	POSITIVE
blid**t**	NEUTER	SINGULAR			PREDICATIVE	POSITIVE
blid**e**		SINGULAR	DEFINITE		ATTRIBUTIVE	POSITIVE
blid**e**		PLURAL				POSITIVE
blid**ere**						COMPARATIVE
blid**este**					ATTRIBUTIVE	SUPERLATIVE
blid**est**					PREDICATIVE	SUPERLATIVE
blid**t**				TRANSADVERBIAL		POSITIVE
blid**ere**				TRANSADVERBIAL		COMPARATIVE
blid**est**				TRANSADVERBIAL		SUPERLATIVE

Table 11. Adjective declension

Transcategorization
This feature relates the word forms which are derived directly from the adjective and function as adverbs to the inflectional paradigm. True (or fully lexicalized) adverbs also exist in parallel, these are provided with the lexical category 'adverb'.

Ex.:

En lovligt varslet konflikt	'lawfully, duly, legally'	(Lit: A legally notified conflict)
En lovlig stor opgave	'rather (too), a bit (too)'	(Lit: A rather big task)

Function
Although the function is a mainly syntactic feature, it is necessary to distinguish the two functions because the use of the particular inflected forms in positive and superlative depends on the function of the adjective.

Comparison
In Danish, comparison by means of suffixes is part of the inflectional paradigm, analytic (or also called periphrastic) comparison forms are not part of the inflection. Further, for semantic reasons, some adjectives cannot be compared at all, e.g. *daglig* 'daily, everyday'.

For all exceptions, etc. please consult the Danish grammar of Allen et al (1995, cf. Reference list).

4.5.3 Verbs

The lexical category of verbs comprises two subcategories: main and medial ('medial' is currently used as a label for deponent verbs, viz. a verb with a passive morphology but functioning as an active verb).
The subcategory main (*adoptere* 'adopt') is by far the most common and largest one.
The subcategory medial comprises only a very few items, such as *lykkes* 'succeed'.

The morphological unit of a verb is its basic form, i.e. the infinitive.

For verbs, the category specific features are as follow: tense, mood and voice.

Transcategorization
This feature relates the word forms which are derived directly from the verb and function as adjectives (viz. present and past participle forms) or nouns (viz. the gerund form).

The table below shows the inflection features of a main verb having a full paradigm.

WORD FORM	GENDER	NUMBER	DEFINITENESS	TENSE	MOOD	VOICE	TRANSCAT
adoptere					INFINITIVE	ACTIVE	
adopteres					INFINITIVE	PASSIVE	
adopterer				PRESENT	INDICATIVE	ACTIVE	
adopteres				PRESENT	INDICATIVE	PASSIVE	
adopterede				PAST	INDICATIVE	ACTIVE	
adopteredes				PAST	INDICATIVE	PASSIVE	
adopter					IMPERATIVE		
adopterende				PRESENT	PARTICIPLE		

adopteret				PAST	PARTICIPLE		
adopteren	COMMON	SINGULAR	UNMARKED		GERUND[1]		TRANSNOMINAL
adopterende	UNMARKED	UNMARKED	UNMARKED	PRESENT	PARTICIPLE		TRANSADJECTIVAL
adopteret	COMMON	SINGULAR	INDEFINITE	PAST	PARTICIPLE		TRANSADJECTIVAL
adopteret	NEUTER	SINGULAR	INDEFINITE	PAST	PARTICIPLE		TRANSADJECTIVAL
adopterede	UNMARKED	SINGULAR	DEFINITE	PAST	PARTICIPLE		TRANSADJECTIVAL
adopterede	UNMARKED	PLURAL	UNMARKED	PAST	PARTICIPLE		TRANSADJECTIVAL

Tabel 12. Attributes and possible values illustrated by a verb with a full inflectional pattern.

5 Frequency information in STO

STO has been provided with frequency information from the two large Danish corpora Korpus 2000 and Korpus 90, comprising texts from 1998-2002 and 1988-1992 respectively. Each corpus consists of 28 mill. words.
The corpora have been automatically annotated with POS-tags using a Brill tagger trained with the PAROLE tag set (see http://korpus.dsl.dk/paroledoc_dk.pdf for more info (in Danish)).

The frequency information consists of four frequency numbers for each word form since the part-of-speech frequency as well as the word form frequency from both corpora is shown.
e.g. håndtryk; NCN_indef_pl;**4**;**112**;7;106
håndtryk; NCN_indef_sg;**80**;**112**;97;106

The first number is the **POS frequency** from Korpus90 which specifies the number of times the word form appears in the corpus with exactly that part of speech. Here it shows that 'håndtryk' appears with the NCN_indef_pl (common noun, neuter, indefinite, plural) tag **4** times and with the NCN_indef_sg (common noun, neuter, indefinite, singular) **80** times.

The second number is the **WF frequency** from Korpus90 that specifies the total number of times that the word form appears in the corpus regardless of the POS tags. Here it shows that the word form 'håndtryk' appears in Korpus90 **112** times. Since the POS frequency in total for both word forms is only 84, it shows that for 28 of the appearances of the word form it has not been possible automatically to assign one of the two right POS tags. So the POS frequency in such cases will be biased.

The two last numbers are POS frequency and WF frequency from Korpus2000 and they illustrate that only 2 appearances have not automatically been assigned one of the two correct tags.
If a word form has not been found in the corpus at all, the frequency numbers are 0. The number -1 has been assigned to POS frequencies in cases where the POS tagger has not assigned the correct POS tag to the word form, e.g.

eskimoisk;A_com_sg_indef_att;-1;11;-1;3
eskimoisk;A_com_sg_unm_pr;-1;11;-1;3
eskimoisk;A_neut_sg_indef_att;-1;11;-1;3
eskimoisk;A_neut_sg_unm_pr;-1;11;-1;3

[1] The English term '*gerund*' is used commonly for the –ing derivative, which is used as a noun. Thus, this term is also used in the present documentation for substantivized verb forms (which is not identical with the meaning of the Danish term '*gerundium*').

eskimoisk;A_tadv_pos;-1;11;-1;3

Due to the detailed and complex tags of this word form, the automatic tagger has not been able to determine which tag is correct for each occurrence of the word form. So for this word form only the WF frequency can be used.
See appendix F for more details on the frequency information file.

6 Literature

Andreasen, Troels, P.A. Jensen, J.F. Nilsson, P. Paggio, B.S. Pedersen, H.E.Thomsen (2004). *Content-based text querying with ontological descriptors*, in: Database and Knowledge Engineering Journal no. 48: pp. 199-219, Elsevier Science B.V., Holland.

Allan, Robin, Ph. Holmes & T. Lundskær-Nielsen (1995). *Danish - A Comprehensive Grammar*, Routledge, London and New York.

Asmussen, Jørg: (2002). *Korpus 2000*, in: Nydanske Sprogstudier 30, p. 27-38, København.

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

Bresnan, Joan (2001). *Lexical Functional Syntax*, Blackwell Textbooks in Linguistics, Blackwell Publishers, Mass. USA.

Braasch, Anna & O. Norling-Christensen (1997). *En trækbaseret beskrivelse af dansk bøjningsmorfologi*, in: Datalingvistisk Forenings årsmøde 1996, Proceedings.

Braasch, Anna, B. Maegaard, B. Pedersen (1998). *En stor dansk sprogteknologisk ordbog - et nationalt projekt*. I: Datalingvistisk Forenings årsmøde 1997, Proceedings, HHS, Kolding.

Braasch, Anna & S. Olsen (2000). *Formalised Representation of Collocations in a Danish Computational Lexicon,* in U. Heid & al., (eds.) Proceedings of the Ninth EURALEX Congress. Stuttgart. p.475-488. (http://cst.dk/sto/referencer/collocations.html)

Braasch, Anna (2004). *A Health Corpus Selected and Downloaded from the Web - Is it Healthy Enough?,* in: S. Vessier & G. Williams (eds.) Proceedings of the XI. EURALEX International Congress, Lorient. Vol. I. pp. 71-79.

Diderichsen, Paul (1987). *Elementær Dansk Grammatik*. Gyldendal. København (3. udg. 9. oplag)

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.

Grimshaw, Jane B. (1990). *Argument Structure*, MIT Press, Cambridge, Mass., US.

Hansen, Aage (1967). *Moderne dansk grammatik,* Grafisk Forlag, København.

Harder, Peter, L. Heltoft & O. Nedergaard Thomsen (1996). *Danish directional adverbs, content syntax and complex predicates: A case for host and co-predicates*, in: E. Engberg-Pedersen et al. (eds.) Content, Expression and Structure. Studies in Danish Functional Grammar pp.159-198. John Benjamins, Amsterdam.

Helbig, Gerhard & W. Schenkel (1980). *Wörterbuch zur Valenz und Distribution deutscher Verben.* Bibliographisches Institut*, Berlin.*

Herslund, Michael & F. Sørensen. (1985). *De franske verber 1. En valensgrammatisk fremstilling. Verbernes syntaks*. Romansk Institut, Københavns Universitet.

Jørgensen, Lise D. and Kirchmeier-Andersen (eds.)(1991). *Eurotra Ordbogsmanual*, Eurotra-DK, Copenhagen.

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

Kilgarriff, Adam (1998). *'SENSEVAL: An Exercise in Evaluating Word Sense Disambiguation Programs,* University of Brighton.

Kilgarriff, Adam & M. Rundell (2002). *Lexical Profiling Software and its Lexicographic Applications – A Case Study,* in: Proceedings of the Tenth EURALEX Congress, Copenhagen. (pp. 807-818.) Center for Sprogteknologi.

Kirchmeier-Andersen, Sabine (1997). *Verbal and Nominal Valency – Sense Distinctions and Inheritance*. In Van Durme, K. (ed) The Valency of Nouns, Odense Working Papers in Language and Communication, no.15, pp.59-86. Odense.

Kirchmeier-Andersen, Sabine (2002). *Dansk korpusbaseret forskning*, in: Nydanske Sprogstudier 30, p. 11-26, København.

Koskenniemi, Kimmo, (1983). *Two-level Morphology: A General Computational Model for Word-form Recognition and Production,* Helsinki.

Navarretta, Costanza (1997). *Danish Syntax Lexicon: Coding Manual for verbs,* upubliceret LE-PAROLE rapport, Center for Sprogteknologi, København

Navarretta, Costanza (1998): *Danish Lexicon: Coding Manual for Adjectives*, upubliceret LE-PAROLE rapport, Center for Sprogteknologi, København

Norling-Christensen, Ole. & Asmussen, J. (1998). *The Corpus of the Danish Dictionary*, Lexikos & Stellenbosch.

Olsen, Sussi (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.

Pollard Carl and I. A. Sag (1987). *Information-Based Syntax and Semantics*, Vol. I: Fundamentals, Center for the Study of Language and Information.

Pollard, C. & I. Sag (1994). *Head-Driven Phrase-Structure Grammar,* The University of Chicago Press, Chicago & London.

Pustejovsky, James (1995). *The Generative Lexicon*, The MIT Press.

Scheuer, Jann (1995). *Tryk på Danske Verber*, RASK Supplement, Vol. 4, Odense Universitetesforlag, Odense.

Schøsler Lene and K. Van Durme (1996). *The Odense Valency Dictionary: An introduction*, Odense Working Papers in Language and Communication, No.13, sept.

Sinclair, John (1991). *Corpus, Concordance, Collocations*, Oxford University Press.

Somers, Harold L. (1997). *Valency and Case in Computational Linguistics*, Edinburgh University Press.

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

Underwood, Nancy L. (ed.) (2000). *The Linda Manual – Typed Feature-based Specifications for a Core Grammar of Danish*, CST Working Papers.

Ørsnes, Bjarne and P. Paggio (1994). *Maskinoversættelse af Substantivkomposita*, in: Baron, I. (ed.) NORDLEX-Projektet: Sammensatte substantiver i dansk, vol. 20 of LAMBDA, pp 135-57. København.

Ørsnes, Bjarne (1995). *The Derivation and Compounding of Complex Event Nominals in Modern Danish - an HPSG Approach with an Implementation in Prolog,* University of Copenhagen.

Reports

EAGLES (1996). *Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora. A Common Proposal and Applications to European Languages*, ILC, Pisa, May 1996.

LINDA (2000). *The LINDA Manual. Typed Feature-based Specifications for a Core Grammar of Danish,* Underwood, N. (ed.), C. Povlsen, P. Paggio, A. Neville, B. Sandford Pedersen, L. Damsgaard Jørgensen, B. Ørsnæs & A. Braasch. Working Papers. Center for Sprogteknologi.

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

Dictionaries:

NDO1999. *Politikens Nudansk Ordbog med etymologi 1999,* 1. udgave, 1.oplag. Politikens Forlag A/S, København (elektronisk udgave).

RO1986. *Retskrivningsordbogen 1986,* Dansk Sprognævn, København

RO1996. *Retskrivningsordbogen 1996,* 2. udgave, Dansk Sprognævn, Aschehoug, København.

RO2001. *Retskrivningsordbogen 2001,* 3. udgave, version 1., Dansk Sprognævn, København (electronic version).

Appendix A

Specifications for morphology export from the STO lexicon

Nouns

Type of information	**Explanation and/or examples**	**Values allowed**
Spelling	The word in canonical form, e.g. *hæfte.* If a word can be inflected in different ways, the *spelling* will appear in two or more consecutive lines followed by the inflected word forms.	
Mu_id	Morphological unit. If a word has more than 1 spelling, these are connected in one MU. The MU HÆFTE_1 covers the spellings *hæfte* and *hefte* meaning 'booklet'. HÆFTE_2 covers the noun *hæfte* and *hefte* meaning 'penalty'. HÆFTE_3 cover the verb *hæfte* og *hefte*.	
Lexcat	Part of speech	NOUN
Sublexcat	Subdivision of the part of speech into subcategories, viz. *common* nouns and *proper* names for nouns.	COMMON PROPER
RO_A	RO-approved States whether the lemma is approved by Retskrivningsordbogen 2001.	YES NO
Origin	States whether a lemma belongs to the general language vocabulary or to a language for specific purposes. Lemmas from general language are marked PAROLE or DDO, depending on the time they were selected. Lemmas from language for specific purposes are labelled with the name of the corpus from which they were selected.	DDO EDB-KORPUS FINANSKORPUS FORVALT-KORPUS H_OG_E-KORPUS MILJØ-KORPUS ONTOQUERY PAROLE SUNDHEDSKORPUS
Decomp	Only used for noun+noun compounds which are decomposed into their two immediate noun components and the joining element between them, if any.	
Fuge	Joining element. Part of the nouns have information on what sign or character, if any has to be removed	

	and/or added when the lemma is the first component of a compound. Letters in square brackets mark the part that has to be removed before the joining element is added, e.g. [e]s, arbejd*e* → arbejdsmand, [] papir → papirklip (viz. nothing removed, nothing added.) Some words have more than one possible joining element, these are separated by a slash, '/'.	
Gender	Gender of nouns. For nouns having plural form only, it is usually difficult to determine the gender. These nouns have the value *unmarked*.	COMMON NEUTER UNMARKED
Ginp	Graphical Inflectional Paradigm. A name for the specific paradigm that reflects the inflection of the lemma. *MFG0662*	
indef_sg	Indefinite, singular form of the lemma *lampe*	
indef_sg_gen	Indefinite, singular, genitive form Until the release of RO2001 various genitive suffixes were allowed, for words ending in –s, -x and –z. Now only the ending –' is approved by RO. In order to be able to recognize formerly used word forms in texts, STO still includes these forms marking them with an *. *lampes* *hus' /*huses /*hus's*	
def_sg	Definite singular form *lampen*	
def_sg_gen	Definite singular, genitive form *lampens*	
indef_pl	Indefinite, plural form *lamper*	
indef_pl_gen	Indefinite, plural, genitive form *lampers*	
def_pl	Definite, plural form *lamperne*	
def_pl_gen	Definite, plural, genitive form *lampernes*	
unm_sg	Mostly proper nouns that do not have inflection as indefinite/definite *Venus*	
unm_sg_gen	Mostly proper nouns that do not have inflection as indefinite/definite, genitive form	

	Venus'	
unm_unm	Indeclinable noun *dart*	

Appendix B

Specifications for morphology export from the STO lexicon

Verbs

Type of information	Explanation and/or examples	Values allowed
Spelling	The word in canonical form, e.g. *hæfte*. If a word can be inflected in different ways, the *spelling* will appear in two or more consecutive lines followed by the inflected word forms.	
Mu_id	Morphological unit. If a word has more than 1 spelling, these are connected in one MU. The MU HÆFTE_1 covers the spellings *hæfte* and *hefte* meaning 'booklet'.. HÆFTE_2 covers the noun *hæfte* and *hefte* meaning 'penalty'. HÆFTE_3 cover the verb *hæfte* og *hefte*.	
Lexcat	Part of speech	VERB
Sublexcat	Subdivision of the part of speech into subcategories, viz. into *main* and *medial* (deponent) verbs.	MAIN MEDIAL
RO_A	RO-approved States whether the lemma is approved by Retskrivningsordbogen 2001	YES NO
Origin	States whether a lemma belongs to the general language vocabulary or to a language for specific purposes. Lemmas from general language are marked PAROLE or DDO, depending on the time they were selected. Lemmas from language for specific purposes are labelled with the name of the corpus from which they were selected.	DDO EDB-KORPUS FINANSKORPUS FORVALT-KORPUS H_OG_E-KORPUS MILJØ-KORPUS ONTOQUERY PAROLE SUNDHEDSKORPUS
Ginp	Graphical Inflectional Paradigm. A name for the specific paradigm that reflects the inflection of the lemma, e.g. *MFG0662*	
inf_act	Infinitive active form of the verb *adoptere*	
inf_pas	Infinitive passive form *adopteres*	

pres_act	Present active form *adopterer*	
pres_pas	Present passive form *adopteres*	
past_act	Past active form adopterede	
past_pas	Past passive form adopteredes	
imp	Imperative form adopter	
pres_part	Present participle form *adopterende*	
perf_part	Past participle form *adopteret*	
nom	Nominalization of the verb *adopteren*	
pres_part_adj	Present participle form used as an adjective *adopterende*	
perf_part_adj_comm_sg_indef	Past participle form used as an adjective; common, singular, indefinite *adopteret*	
perf_part_adj_neut_sg_indef	Past participle form used as an adjective; neuter, singular, indefinite *adopteret*	
perf_part_adj_unm_sg_def	Past participle used as an adjective. Gender unmarked, singular, definite *adopterede*	
perf_part_adj_unm_pl_unm	Past participle used as an adjective. Gender unmarked, plural, definiteness unmarked *adopterede*	

Appendix C

Specifications of morphology export from the STO lexicon

Adjectives

Type of information	Explanation and/or examples	Values allowed
Spelling	The word in canonical form, e.g. *hæfte*. If a word can be inflected in different ways, the *spelling* will appear in two or more consecutive lines followed by the inflected word forms.	
Mu_id	Morphological unit. If a word has more than 1 spelling, these are connected in one MU. The MU HÆFTE_1 covers the spellings *hæfte* and *hefte* meaning 'booklet'.. HÆFTE_2 covers the noun *hæfte* and *hefte* meaning 'penalty'. HÆFTE_3 cover the verb *hæfte* og *hefte*.	
Lexcat	Part of speech	ADJECTIVE
Sublexcat	Subdivision of part of speech into subcategories. Adjectives are subdivided into normal, cardinal and ordinal.	CARDINAL NORMAL ORDINAL
RO_A	RO-approved Tells whether the lemma is approved by the Retskrivningsordbogen 2001	YES NO
Origin	States whether a lemma belongs to the general language vocabulary or to a language for specific purposes. Lemmas from general language are marked PAROLE or DDO, depending on the time they were selected. Lemmas from language for specific purposes are labelled with the name of the corpus from which they were selected.	DDO EDB-KORPUS FINANSKORPUS FORVALT-KORPUS H_OG_E-KORPUS MILJØ-KORPUS ONTOQUERY PAROLE SUNDHEDSKORPUS
Ginp	Graphical Inflectional Paradigm. A name for the specific paradigm that reflects the inflection of the lemma. *MFG0662*	
com_sg_indef_att	Common, singular, indefinite, attributive, positive form *blid*	

neut_sg_indef_att	Neuter, singular, indefinite, attributive, positive form *blidt*	
unm_sg_def_att	Gender unmarked, singular, definite, attributive, positive form *blide*	
com_sg_unm_pr	Common, singular, definiteness unmarked, predicative, positive form *blid*	
neut_sg_unm_pr	Neuter, singular, definiteness unmarked, predicative, positive form *blidt*	
unm_pl_unm_unm	Gender unmarked, plural, definiteness unmarked, function unmarked, positive form *blide, atten, tredje*	
comp	Comparative form *blidere*	
att_sup	Attributive, superlative form *blideste*	
pre_sup	Predicative, superlative form *blidest*	
tadv_pos	Transadverbial (adjective used as an adverb) form *blidt*	
tadv_comp	Transadverbial (adjective used as an adverb), comparative form *blidere*	
tadv_sup	Transadverbial (adjective used as an adverb), superlative form *blidest*	

Appendix D

Specifications for morphology export from the STO lexicon

Other parts of speech

Type of information	Explanation and/or examples	Values allowed
Spelling	The word in canonical form, e.g. *hæfte*. If a word can be inflected in different ways, the *spelling* will appear in two or more consecutive lines followed by the inflected word forms.	
Mu_id	Morphological unit. If a word has more than 1 spelling, these are connected in one MU. The MU HÆFTE_1 covers the spellings *hæfte* and *hefte* meaning 'booklet'.. HÆFTE_2 covers the noun *hæfte* and *hefte* meaning 'penalty'. HÆFTE_3 cover the verb *hæfte* og *hefte*.	
Lexcat	Part of speech Adpositions concern in Danish prepositions only. Unique are words like *som, der, at* which cannot clearly be classified as any other part of speech..	ADPOSITION ADVERB CONJUNCTION INTERJECTION UNIQUE
Sublexcat	Subdivision of part of speech into subcategories or minor groups. All adverbs have the sub-lexcat general. All adpositions have the sub-lexcat preposition.	ADV: GENERAL ADP: PREPOSITION
RO_A	RO-approved Tells whether the lemma is approved by the Retskrivningsordbogen 2001	YES NO
Origin	States whether a lemma belongs to the general language vocabulary or to a language for specific purposes. Lemmas from general language are marked PAROLE or DDO, depending on the time they were selected. Lemmas from language for specific purposes are labelled with the name of the corpus from which they were selected.	DDO EDB-KORPUS FINANSKORPUS FORVALT-KORPUS H_OG_E-KORPUS MILJØ-KORPUS ONTOQUERY PAROLE SUNDHEDSKORPUS
Ginp	Graphical Inflectional Paradigm. A name for the specific paradigm that	

	reflects the inflection of the lemma, *MFG0662*	

Appendix E

Specifications for morphology export from the STO lexicon

Pronouns

Type of information	Explanation and/or examples	Values allowed
Spelling	The word in canonical form, e.g. *hæfte*. If a word can be inflected in different ways, the *spelling* will appear in two or more consecutive lines followed by the inflected word forms.	
Mu_id	Morphological unit. If a word has more than 1 spelling, these are connected in one MU. The MU HÆFTE_1 covers the spellings *hæfte* and *hefte* meaning 'booklet'.. HÆFTE_2 covers the noun *hæfte* and *hefte* meaning 'penalty'. HÆFTE_3 cover the verb *hæfte* og *hefte*.	
Lexcat	Part of speech	PRONOUN
Sublexcat	Subdivision of part of speech into subcategories.	DEMONSTRATIVE INDEFINITE INTERROGATIVE PERSONAL POSSESSIVE RECIPROCAL
RO_A	RO-approved Tells whether the lemma is approved by the Retskrivningsordbogen 2001	YES NO
Origin	States whether a lemma belongs to the general language vocabulary or to a language for specific purposes. Lemmas from general language are marked PAROLE or DDO, depending on the time they were selected, lemmas from language for specific purposes are labelled with the name of the corpus from which they were selected.	DDO EDB-KORPUS FINANSKORPUS FORVALT-KORPUS H_OG_E-KORPUS MILJØ-KORPUS ONTOQUERY PAROLE SUNDHEDSKORPUS
Ginp	Graphical Inflectional Paradigm. A name for the specific paradigm that reflects the inflection of the lemma. *MFG0662*	

Personal Pronouns

pron_pers_nom	Personal pronoun, nominative *jeg, du, han, hun, det, vi, I, de, De*	
pron_pers_unm	Personal pronoun, case unmarked *mig, dig, ham, hende, det, os, jer, dem, Dem*	
pron_pers_3_unm_unm_unm_ref	Personal pronoun, 3. person, number unmarked, gender unmarked, case unmarked, reflexive *sig*	

Possessive pronouns

pron_poss_sg_com	Possessive pronoun, singular, common *min, din, sin, , vor,*	
pron_poss_sg_neu	Possessive pronoun, singular, neuter, *mit, dit, sit, , vort,*	
pron_poss_pl_unm	Possessive pronoun, plural, gender unmarked *mine, dine, sine, , vore,*	
pron_poss_unm_unm	Possessive pronoun, number unmarked, gender unmarked *hans, hendes, vores, jeres, deres, Deres*	

Demonstrative pronouns

pron_demon_com_sg_unm	Demonstrative pronoun, common, singular, case unmarked *denne*	
pron_demon_com_sg_gen	Demonstrative pronoun, common, singular, genitive *dennes*	
pron_demon_neu_sg_unm	Demonstrative pronoun, neuter, singular, case unmarked *dette*	
pron_demon_neu_sg_gen	Demonstrative pronoun, neuter, singular, genitive *dettes*	
pron_demon_unm_pl_unm	Demonstrative pronoun, gender unmarked, plural, case unmarked *disse*	
pron_demon_unm_pl_gen	Demonstrative pronoun, gender unmarked, plural, genitive *disses*	
pron_demon_unm_unm_unm	Demonstrative pronoun, gender, number and case	

	unmarked *selv*	

Reciprocal pronouns

pron_rec_unm_pl_unm	Reciprocal pronoun, gender unmarked, plural, case unmarked *hinanden*	
pron_rec_unm_pl_gen	Reciprocal pronoun, gender unmarked, plural, genitive *hinandens*	

Interrogative pronouns

pron_inter_sg	Interrogative pronoun, singular gender and case unmarked *hvad*	
pron_inter_com	Interrogative pronoun, common, number and case unmarked *hvem*	
pron_inter_gen	Interrogative pronoun, genitive, number and gender unmarked *hvis*	
pron_inter_com_sg_unm	Interrogative pronoun, common, singular, case unmarked *hvilken*	
pron_inter_neu_sg_unm	Interrogative pronoun, neuter, singular, case unmarked *hvilket*	
pron_inter_unm.pl.unm.	Interrogative pronoun, plural, gender and case unmarked *hvilke*	

Indefinite pronouns

pron_indef_com_sg_unm	Indefinite pronoun, common, singular, case unmarked *anden*	
pron_indef_com_sg_gen	Indefinite pronoun, common, singular, genitive *andens*	
pron_indef_neu_sg_unm	Indefinite pronoun, neuter, singular, case unmarked *andet*	
pron_indef_neu_sg_gen	Indefinite pronoun, neuter, singular, genitive *andets*	
pron_indef_unm_pl_unm	Indefinite pronoun, plural, gender and case unmarked	

	andre	
pron_indef_unm_pl_gen	Indefinite pronoun, plural, genitive, case unmarked *andres*	
pron_indef_com_nom	Indefinite pronoun, common, nominative, number unmarked	

Appendix F

Specification for morphology export from the STO lexicon

Frequency information

Type of information	Explanation and/or examples	Values allowed
Spelling	The word in canonical form, cf. the different word categories	
Lexcat	Part of speech	
Sublexcat	Subdivision of part of speech into subcategories	e.g. common, proper (nouns) personal, demonstrative etc. (pronouns)
RO_A_gmu	RO_approved lemma Shows whether this lemma is approved by Retskrivningsordbogen 2001	YES NO
RO_A_gmu_ginp	RO_approved inflectional paradigm States whether the inflectional paradigm for this lemma is approved by Retskrivningsordbogen 2001	YES NO
Ginp	Graphical Inflectional Paradigm. The name for the specific paradigm that reflects the inflection of the lemma, e.g. *MFG1023*	
Wordform	The word form found in the corpus.	
Pos	Part_of_speech-tag. The tag that specifies the part of speech and the other morphological features of the word form e.g. *NCN_indef_pl*	
Pos_freq_K90	POS tag frequency in Korpus 90 The number of times the word form appears with that specific POS tag in Korpus 90	
Wf_freq_K90	Word form frequency in Korpus 90 The number of times the word form appears in Korpus 90 regardless of POS-tag.	
Pos_freq_K2000	POS tag frequency in Korpus 2000 The number of times that the word form appears with that specific POS tag in Korpus	

	2000.	
Wf_freq_K2000	Word form frequency in Korpus 2000 The number of times the word form appears in Korpus 2000 regardless of POS-tag.	

PART 2:

DOCUMENTATION OF THE SYNTACTIC LAYER

NOTE: The description of the XML structure of the syntax of STO enclosed here is the former XML structure. The current LMF structure is described in an independent documentation file:
http://cst.ku.dk/sto_ordbase/STO-LMF-syntax-documentation-v1.pdf
However, the other parts of this documentation are still relevant for comprehending the syntax of STO.

Introduction

The acronym **STO** is the abbreviation of the Danish title of the present lexicon, and it stands for '**S**prog**T**eknologisk **O**rdbase', Lexicon for Language Technology Applications.
This acronym is used in the file names of the delivered data material.

The data material is provided as three XML files as follows (size in bytes):

STO_Syntax_1_v1.xml	4437723
STO_Syntax_2_v1.xml	4872856
STO_Syntax_3_v1.xml	4488728

The data files can be validated with the XML Schema which can be found in Appendix 1. (File name: STO_Syntax.xsd, size 21865 bytes).

The present documentation is the second part of the complete Documentation of the STO lexicon, the first part provides a general description of the STO lexicon and describes the Morphological Layer. Part 2 of a Documentation on the STO Syntactic Layer is structured in the following way: Firstly, some general decisions concerning the syntactic layer are described briefly (Chapter 1), secondly, the linguistic principles adopted for the encoding of syntactic features are presented for each word class in detail (Chapter 2), and finally we describe the format of the XML-files (Chapter 3).

The Documentation contains also a literature list, which may be useful for the user not familiar with the linguistic theories mentioned in the document or with the Danish language. For details of the Danish grammar please consult Allen et al. *A Comprehensive Grammar of Danish*' (1995).

1 The syntactic layer

1.1 The contents of the lexicon

The vocabulary provided with a syntactic description is a proper subset (app. 45,000 words) of the whole vocabulary covered in STO (app. 81,300 words). Thereof approximately 34,000 words are selected from the general language corpora of STO on the basis of their frequency. The rest, approx. 11,000 words belong to the domain languages covered. All syntactic information contained in STO is encoded on the basis of corpus evidence, where preference is given to more frequent syntactic patterns in cases, where an exhaustive encoding (including single or sparse occurrences of a single word) was not possible because of the overwhelming number of syntactic constructions. As an example can be mentioned a group of very common verbs (e.g. *komme* (to come), *tage* (to take)) having different syntactic constructions in such numbers that the list including also their infrequent constructions would be too comprehensive.

The methodology adopted in the description of syntactic features of words is mainly based on the valency theory. Main attention is paid to various grammatical structures of words consisting of a dominant word (noun, adjective or verb) and its complements. Thus, the syntactic layer of STO comprises the description of valency patterns for nouns, verbs and adjectives. Adverbs are not provided with syntactic descriptions in this data delivery (Version 1), although the appropriate

structure of XML elements is implemented (cf. Chapter 3 The Data: STO syntax represented as XML elements).

1.2 Linguistic description at the syntactic layer

1.2.1 Basic principles of valency-boundness

The central features in the linguistic description at the syntactic layer concern the valency of words, and the syntactic behaviour of words are described in terms of valency patterns.

The approach to valency adopted in STO is highly inspired by the model of six distinct degrees of valency-binding, developed in Somers (1987). This model is based on the observation that there are different degrees by which constituents depend on the head. Most dependent are the lexically determined integral complements which are typically parts of the predicate e.g. in collocations; thereafter come the obligatory and optional complements. 'Middles'(ibid., p. 27) are on the borderline between complements and modifiers where adjuncts or modifiers are the less dependent.

The encoding strategy for complements and middles vs. adjuncts is based on the above sketched model. Since collocations (and other types of multi-word units) do not form part of STO yet, integral complements are not encoded. All **obligatory** and **optional complements** are encoded while adjuncts are not seen as part of the valency pattern, unless they are frequent.

In general '**middles**' are encoded in cases where they are noticeably frequent in the corpora e.g. *en markedsandel på 20%* (a market share of 20%) and if they are pertinent to the central meaning of the word (e.g. *et mindre afbræk i produktionen* (lit.: a minor break in the production). The treatment of 'middles' varies slightly, dependent on the word class in question. The strategy adopted for the individual word classes is described in the relevant sections.

A few examples below illustrate some prototypical syntactic constructions that are regarded as valency-bound and accordingly, they are described in STO. (The optional complements are in brackets.)

- Verbal constructions
  læse (bøger) (read (books)): divalent (transitive)
  sende et brev (fra København) (til Århus) (send a letter (from Copenhagen) (to Århus)): trivalent
- Nominal constructions
  en kasse æbler (a box of apples): monovalent
  en søster til Peter (a sister of Peter)
- Adjectival construction
  afhængig af indtægter (dependent on income): divalent

Contrary to the above examples, phrases of the type *spise (på restaurant/ om morgenen/ tre gange om dagen)* (eat (at a restaurant/ in the morning/ three times a day)) are not encoded, because the **locative** and **temporal adjuncts** in connection with the verb *spise* (eat) are not pertinent to the meaning of the verb but they are completely arbitrary (free modifiers).

1.2.2 The description of valency

The word is considered to be the syntactic head of the described construction,

- its valency is given in a **valency pattern** by listing and describing the constituents the word subcategorizes for and their characteristics and
- the characteristics of the **head itself** in the current context is given.

A valency pattern is described in a syntactic **description** and encoded in a **syntactic unit**. Thus different syntactic descriptions of the same word are encoded in different syntactic units. Exceptions are valency patterns which are considered to be syntactic alternations. These alternating syntactic descriptions are encoded in the same syntactic unit. Examples of encoded syntactic alternations are the following

- Verbs: dative shift alternation, e.g.
  Jeg giver hende en bog/jeg giver en bog til hende
  (I give her a book/I give a book to her)
- Adjectives: infinitive-clause without/with expletive, e.g.
  Det er attraktivt at bo i København/At bo I København er attraktivt
  (It is attractive to live in Copenhagen/To live in Copenhagen is attractive)

In other words, a valency pattern of a word therefore contains information about

- the arity of that word, viz. how many complements it governs (between 0 and 4)
- the syntactic function of each governed complement (e.g. subject, object)
- the syntactic construction potential of each complement expressed in terms of syntactic categories (e.g. noun phrase, prepositional phrase, clause).

In STO further linguistic details are specified, such as

- the preposition of each governed prepositional phrase and the construction potential of the governed element
- the control type in case of infinite clauses (viz. subject or object control)
- obligatoriness of each governed complement.

1.2.3 Treatment of the control phenomena

As mentioned above, the STO syntax provides information about control phenomena. For example the STO syntax indicates explicitly **subject control.** In the example below, the unexpressed subject of the infinitive *at ringe* (to ring up) in the sentence is the same as the subject of the matrix clause, i.e. *jeg* (I).

Ex.: *Jeg ønskede at ringe til hende* (I wished to ring her up)

Another control type encoded is the **object control**. In the following example, the unexpressed subject of the infinitive *at ringe* in the sentence is the object of the matrix clause, i.e. *Maria*.

Ex.: *Jeg anbefalede Maria at ringe til hende* (I recommended Maria to ring her up)

2 The syntactic description of word classes

2.1 Verbs

This section describes the syntax of verbs. Firstly, some examples illustrate the distinction between complements and adjuncts/free modifiers. Secondly, a list of syntactic functions of the complements is provided and commented. Thirdly, the arity and valency patterns of verbs are described. Fourthly, the treatment phrasal verbs and reflexive verbs are sketched. Finally, the verbal alternations are listed.

2.1.1 Complements vs. adjuncts/free modifiers

According to the general principle (cf. 1.2 Basic principles and description elements), all obligatory and optional complements are encoded, whereas adjuncts are not seen as part of the valency pattern. The strategy adopted for the treatment of middles (Somers, op. cit.) is based on the following observation: these elements are semantically determined by the governing verb, while their syntactic function and construction types may vary.

Middles are encoded in the valency pattern in the following cases:

- If they tend to be obligatory
  e.g. *han bor i Spanien/på landet* (he lives in Spain/ in the countryside
- If they express direction with motion verbs
  e.g. *han gik (ned/op/ud ... til gården)* (he vent down/up/out ... to the yard)
- If they are closely related to the core meaning of the verb and occur significantly in the corpus
  e.g. *jeg bød ham 1000 DKK for jobbet* (I offered him 1000 DKK for the job).

2.1.2 Syntactic functions of complements

The description comprises the following syntactic functions:

- SUBJECT
- OBJECT
- INDIRECT_OBJECT
- PREPOSITIONAL_OBJECT
- OBJECT_PREDICATE
- FORMAL_SUBJECT
- ADVERBIAL
- SUBJECT_PREDICATE

Some of the above listed functions deserve further explanation.
FORMAL_SUBJECT is a place-holder or an impersonal subject occurring in *der* and *det* constructions in POSITION 1, the real SUBJECT occupies POSITION 2 or POSITION 3 depending on the other constituents of the construction. Exceptions are weather verbs which do not refer to a subject.

Ex.: *Der var tre mænd i haven* (There was three men in the garden)
Det interesserer mig om du kommer (lit: It interests me whether you come)
Det regner (lit: It rains; It's raining.)

INDIRECT_OBJECT is used for a group of ditransitive verbs, and it can occur in POSITION 2 or POSITION 3.

Ex.: *Jeg giver hende bogen* (I give her the book)
Jeg giver bogen til hende (I give the book to her)

PREPOSITIONAL_OBJECT refers to an object introduced by a valency-bound preposition, as in:
Ex.: *Vi tror på Gud* (We believe in God)

ADVERBIAL is encoded if it is governed by the verb
Ex.: *Han behandler hende godt* (He treats her well)

2.1.3 The arity of verbs and the numbering of the positions in constructions

The arity feature registers the number of valency-bound semantic arguments occurring in a given construction. This feature may be relevant for human inspection or in the case of semantic encoding, and it is registered in the identifier of the *Description* and *Construction* elements, e.g. Dv3... or Cv2… . The following verb arity types are registered: zero-, mono-, di-, tri- and tetravalent (see the examples below). In the following, the strategy adopted for the counting of arity in verb valency patterns are explained and illustrated. Some elements of particular construction types are not included into the semantic arity number (e.g. the FORMAL_SUBJECT of weather verbs), but they are represented within the syntactic description of the construction by an appropriate POSITION.

FORMAL_SUBJECT is not included into the semantic arity, although represented as a POSITION in the construction
Ex.: *Det regner* (lit: It rains; It's raining). Weather verbs are considered zerovalent.
Der var tre mænd i haven (There were three men in the garden). This is considered to be a monovalent construction.

OBJECT_PREDICATE is not included into the arity because it is regarded as being a part of the OBJECT, although it is represented as a POSITION in the construction. The same applies for accusative constructions with infinitive.
Ex.: *De kaldte ham klog* (lit.: They called him clever)
Jeg hørte ham komme (lit.: I heard him come)
Both examples are considered divalent constructions

SUBJECT_PREDICATE is considered a semantic argument even if it is not realized as such.
Ex.: *Du ser træt ud* (You look tired): this is considered divalent

ADVERBIAL is counted as a complement if it is governed by the verb
Ex.: *Han behandler hende godt* (He treats her well): is considered trivalent.

The valency pattern refers to an individually numbered position for each of the above mentioned elements.

2.1.4 The arity of reflexive verbs and phrasal verbs

A **true reflexive verb** consists of a verb and a reflexive pronoun, which form a unit, thus it is treated as a particular *Self* type (cf. the section 3.1.4 on Self for verbs). The reflexive pronoun is usually semantically empty in case of true reflexivity, and it is often not translated (cf. the first example below). Therefore the pronoun is not encoded as an individual position in the valency pattern but in the *Self* and the verb is considered intransitive.
Ex.: *Han brokkede sig* (He complained)
Peter morede sig (Peter enjoyed himself)

In contrast, in other reflexive constructions the reflexive pronoun does occupy the position of the OBJECT and the construction is considered transitive.

Ex.: *Han vaskede sig* (He washed himself) a reflexive NP

Peter vaskede sin søn og sig selv (Peter washed his son and himself):

divalent (transitive), the OBJECT contains two co-ordinated NP's, whereof the second one (*sig*) is a reflexive pronoun.

In case of **phrasal verb**s, the particle is treated in the same way, i.e. particle verbs are treated as a particular SELF type.The particle is not encoded as an individual position in the valency pattern but belongs to the *Self* element. (See further section 2.1.6 on Phrasal verbs – The treatment of particles).

Ex.: *Fristen løber ud i morgen* (lit.: The respite runs out tomorrow)

2.1.5 Valency pattern types of verbs - An overview

In the following a number of typical examples are listed in order to illustrate each encoded verb arity type. The complement having the SUBJECT function is not explained in the examples below because the SUBJECT is obligatory and it occupies the POSITION 1 in constructions with canonical word order.(The valency-bound complements are underlined, optional complements are in brackets, alternating structures are in square brackets with a slash between them.)

2.1.5.1 Zerovalent

Verb constructions with FORMAL_SUBJECT only

Ex.: *Det sner i dag* (It is snowing today)

2.1.5.2 Monovalent verb constructions

Verb constructions with SUBJECT, incl. also FORMAL_SUBJECT

Ex.: *Endelig dukkede han op* (Finally he showed up)

Isen smelter i solen (The ice melts in the sun)

Peter morede sig i går (Peter enjoyed himself yesterday)

Det forlyder, at han ikke må rejse (It is said that he is not allowed to leave.)

Det nytter ikke at klage (It doesn't help to complain)

Det lønner sig at arbejde (It pays to work)

2.1.5.3 Divalent

Verb constructions with SUBJECT incl. also FORMAL_SUBJECT and OBJECT, INDIRECT_OBJECT or PREPOSITIONAL_OBJECT.

Ex.: *Han læser (en bog)* (He reads (a book)) (object, optional)

Han afprøver programmet (He tests the software) (object)

Peter vaskede sig i går (Peter washed himself yesterday) (object)

Han troede hende (He trusted her) (indirect object)

Jeg frydes over sangen (lit.: I'm delighted over the song) (prepositional object)

Der hører brød til maden (lit.: there belongs bread to the food)

Brødet hører til maden (The bread belongs to the food)

De kaldte ham klog (lit.: They called him clever) (object predicate)

2.1.5.4 Trivalent

Verb constructions with SUBJECT incl. also FORMAL_SUBJECT, OBJECT and INDIRECT_OBJECT or PREPOSITIONAL_OBJECT, ADVERBIAL…

Ex.: *Marie gav bogen til drengen* (Marie gave the book to the boy)

Marie gav drengen en bog (Marie gave the boy a book)
Jeg fraråder dig at rejse (I dissuade you from leaving)
Han sladrede (til politiet) (om flugten) (He babbled (to the police) (about the flight)
Hun rejser (fra København) (til London) (She travels (from Copenhagen) (to London))

2.1.5.5 Tetravalent

Verb constructions with SUBJECT incl. also FORMAL_SUBJECT, OBJECT and INDIRECT_OBJECT or PREPOSITIONAL_OBJECT(S), ADVERBIAL…)
Ex.: *Firmaet importerer kaffebønner (fra Afrika) (til de nordiske lande)*

2.1.6 Phrasal verbs - Treatment of particles

Phrasal verbs consist of a verb and a particle which is encoded in the *Self* for the verb in question (cf. the section about Self elements), as mentioned before. A number of verbs, mainly expressing motion, can combine with various local or directional particles. In such cases a general value **'dir'** and/or **'loc'** is encoded.
Ex.: *Han gik ud/op/ned på taget/i gården*
(lit.: He went out/up/down [dir] to the roof/ in the courtyard [loc])

Note: The particular directional particle *hen* is not member of the 'dir' particle group because it usually cannot stand alone but combines with an obligatory adverbial. (It is not rendered in English, cf. the example translated below.)
Ex.: *Han gik hen til huset* (He went to the house)

In order to identify phrasal verbs, the method of the so-called 'loss of stress test' is employed (cf. Scheuer 1995 and Harder, Heltoft & Thomsen 1996). 'Loss of stress' is a term of phonetics designating the phenomenon that phrasal verbs have unity stress, i.e. the verb looses its individual stress. Verb + particle co-occurrences without a loss of stress are not encoded as particle verbs.
Ex.: *Han blev 'væk* ((lit.: He stayed away) (He disappeared))
The ' marks the stress, here only the particle is stressed, vs.
Han 'blev 'væk ((lit.: He stayed away) (He didn't come))
In this latter example, both the verb and the particle are stressed.

Note: A number of verbs allow for a preposition without an NP (viz. a particle) and also for a prepositional construction with an NP. In such cases both constructions are encoded.
Ex.: *Han stod af* (He got off/ he opted out)) which is a phrasal verb vs.
Han stod af bussen (He got off the bus).

2.1.7 Treatment of prepositions

In case of locative verbs like *bo* (live) various **locative** prepositions may occur. A general value **'loc'** stands for *i* (in), *på* (on, in), *ved* (at, by), *under* (under), *over* (over), *foran* (before, a head of), *bag* (behind), *bag ved* (behind). An identical encoding principle is adopted in case of motion verbs, where the prepositional phrase is preceded by a directional particle (cf. above).
Ex.: *Vi bor på en ø/i København/ved skoven* (We live on an island/in Copenhagen/ by the forest)
Note: the **directional** prepositions *til* (to), *fra* (from) and *over* (over) are spelled out.

2.1.8 Syntactic units and verb alternations

As has been briefly mentioned in the introductory section on basic principles and description elements, syntactic alternations are encoded in the same syntactic unit. For verbs, the set of alternations encoded as such is limited to the following four (for the sake of clarity, we include the description identifiers in the examples):

- Dative alternations:
  Dv3indNN - *jeg giver hende bogen* (I give her the book) and
  Dv3NPind - *jeg giver en bog til hende* (I give the book to her)

- 'There' alternations
  Dv1 – *en anden verden eksisterer* (another world exists) and
  Dv1exderN – *der eksisterer en anden verden* (there exists another world)

- Reciprocal alternations:
  Dv1 – *vi samarbejder* (we collaborate) and
  Dv2P-med – *jeg samarbejder med dig* (I collaborate with you)

- Reflexive alternations (NB: not for true reflexives)**:**
  Dv3xindNN-ind - *jeg bilder ham noget ind* (I made him believe something) and
  Dv3xindrefNN-ind – *jeg bilder mig noget ind* (I made me believe something)

2.2 Nouns

This section describes the syntax of nouns, the following main topics are discussed. Firstly, the valency of nouns and the distinction between complements and adjuncts/free modifiers and the adopted encoding strategy is explained. Secondly, a commented list of syntactic functions of the complements is provided. Thirdly, the noun complementation types and particular elements of the valency patterns are explained and exemplified.

2.2.1 The valency of nouns

The general principle (as mentioned in Section 1.2) is adopted as a basis for the syntactic description of nouns. However, it is far more difficult in case of nouns to differentiate between 'middles' and modifiers. The strategy for encoding of 'middles' is based on the fact that computational language processing tasks (e.g. recognition of syntactic structures parsing) benefit from an encoding which is broadened to include this type too. Middles are encoded if they were significantly frequent in the corpus and if they were pertinent to the central meaning of the word. The latter has a particular relevance for the treatment of the domain-related vocabulary.

Ex.: *et mindre afbræk i produktionen* (lit.: a minor break in the production)
en markedsandel på 20% (a market share of 20%) vs.
en bog (på 1 kg) (a book (of 1 kg)) where the quantification does not form a part of the central meaning of the noun, thus it is not encoded.

Furthermore, almost every complement of nouns is optional, thus obligatoriness is not a useful criterion for the identification of valency bound complements. Basically, a noun describes an entity (and not a relation), consequently the vast majority of simple (concrete and abstract) nouns are said not to subcategorise for any complement. On the other hand, nouns derived from verbs or adjectives

often take over the valency of the word from which they are derived. (For a further discussion on the valency of nouns please see Grimshaw (1990).)

In the following, a list of valency bearing noun types is given:

Simple, concrete nouns

- Family relationship
  Ex.: *Peters søster/søster til Peter* (Peter's sister/the sister of Peter)
- Pictures, semiotic artifacts
  Ex.: *billedet af Mona Lisa* (the picture of Mona Lisa)
  bogen om landbruget (the book about agriculture)
- Direction nouns
  Ex.: *stien fra huset til stranden* (the path from the house to the beach)

Simple, abstract nouns

- Property
  Ex.: *husets farve* (the colour of the house)
- Mass
  Ex.: *flertallet af indbyggerne* (the majority of the inhabitants)

Deadjectival nouns

These nouns inherit at least one valency-bound complement of the adjective from which they are derived, viz. the external argument of the adjective (cf. Section 2.3.3). In some cases, also the internal argument is inherited.

- External argument
  Ex.: *træets højde* (the height of the tree) from *træet er højt* (the tree is high)
- External and internal argument
  Ex.: *farderens godhed mod dyr* (lit.: the goodness of the father to animals) from
  Faderen er god mod dyr (lit.: the father is good to animals)

Deverbal nouns

These nouns may inherit none, one or more valency-bound complements of the verb from which they are derived, depending on the type of nominalization (cf. Kirchmeier-Andersen (1997:59-86). In STO the encoding strategy is based on the distinction between 'process (or event) nominals' and 'result nominals' as stipulated in Grimshaw (1990, Chaper 3).

- Process nouns have a dynamic semantic content and thus they allow for realization of one or more valency bound complements
  Ex.: *Lægens behandling af patienten for lungebetændelse varede længe*
  (Lit.: The doctor's treatment of the patient for pneumonia took a long time)
- Result nouns have a static semantic content and they are avalent i.e. they do not inherit any complements of the verb
  Ex.: *Bygningen findes i byen* (The building is in the city): result reading
- A large number of deverbal nouns has both a process and a result reading as well
  Ex.: *Bygningen af hovedkvarteret varede flere år*
  (The building of the head quarter took several years): process reading.

Compound nouns

Compound nouns in Danish are written as a single word without blanks between its elements (cf. the Section on morphology). In certain cases, the first element of the compound may be an

incorporated complement of the second element (a), but this is by far not always the case. Even though the first element is not an incorporated complement (b), compound nouns trend to have a less rich complementation than simple nouns. The encoding is also in this case based on corpus evidence.

Ex.: a. *program (til tegning)* -> *tegneprogram*
(lit.: program (for drawing) -> drawing-program) vs.
b. *standardprogram (til tegning)* (a standard program (for drawing))

2.2.2 Optionality of complements and encoding strategy

Generally, complements of nouns are optional, there are only a very few nouns which cannot occur without complements. Obligatory complements are thus explicitly marked in the syntactic description.

Ex.: *kvindernes enemærker* (women's preserves)

Because of the general optionality of complements, it is very often impossible to differentiate between the result reading of a noun and its process reading in an elliptic construction (i.e. where the complements are omitted). Thus, in order to avoid unnecessary quasi-duplicates in the database and over-generation in computational language processing, a generalisation strategy of providing 'broad descriptions' is adopted in the following cases:

- Nouns having both process and result reading are provided with one single syntactic description covering both readings
  Ex.: (*Kommunens) administration (af klausulerne)*
  (lit: (the municipality's) administration (of the clauses)), process reading
  Administrationen har engelsk som arbejdssprog
  (The administration has English as its working language), result reading.
- Nouns which subcategorise for various combinations of optional complements are provided with one single description, viz. the broadest possible complementation pattern, even though a construction with all complements realised is not observed within the corpus.
  Ex.: a. *(deres) flytning (til England)* ((their) moving (to England))
  b. *flytning af arbejdspladser fra København*
  ((moving (of working places) (from Copenhagen))
  c. *flytning (af hæren) (fra Århus) (til Krarup)*
  ((moving (of the army) (from Århus) (to Krarup)

2.2.3 Syntactic functions of complements

- **REL_GEN denotes a NP in genitive (relational genitive), from which, in case of deverbal or deadjectival nouns, the subject of the verb or adjective the noun is derived.**
  Ex.: *Hansens erkendelse af fejlen* (Hansen's acknowledge of the mistake)
- PREPOSITIONAL_OBJECT
  Ex.: *Columbus opdagelse af Amerika* (Columbus' discovery of America)
- CLAUSCOMP denotes clausal complements (that- or infinitive clauses)
  Ex.: *Den metode at anvende datamater til forskning* (the method of using computers in research)
- SPEC_N denotes a nominal complement of mass entities
  Ex.: *en kasse æbler* (a box of apples)

2.2.4 The Self element for nouns

The *Self* for nouns contains one element only, *Definite_Suffix_Allowed*, which can take the values YES or NO. This feature concerns the definiteness of nominal complements (viz. the use of clitic). The value is NOT in case of a nominal complement of a mass entity, in all other cases YES (which is seen as the default value.)
Ex.: *en kasse æbler* (a box of apples)

2.2.5 Valency frames of nouns – An overview

In the following, the various noun complementation types are described and for illustration purposes, a few examples are provided with a detailed syntactic description.
Noun patterns may contain three complement types: genitive, noun and prepositional phrase. A list of prototypical examples illustrates the encoding of noun valency in STO.

Genitive or an alternating prepositional phrase

Simple, concrete nouns are considered monovalent nouns

- Family relationship
  Ex.: a.) *en søster til Peter* (a sister of Peter) PREPOSITIONAL_OBJECT
  (Position: 1; Function: PREPOSITIONAL_OBJECT; Optional:YES; Syntactic_Label: PP; Introducer: TIL; Clause_Type: NULL; NP_Type: N; Control: NOCONTROL; Coref: NOCOREF)
  b.) *Peters søster* (Peter's sister) REL_GEN
  (Position:1; Function: REL_GEN; Optional: YES; Syntactic_Label: NP; Casus: GENITIVE)
  The subgroups below are very similar to the Family relationship subgroup
- Picture nouns
  Ex.: *Mona Lisas billede* (Mona Lisa's picture) REL_GEN
  billedet af Mona Lisa (the picture of Mona Lisa) PREPOSITIONAL_OBJECT
- Professions
  Ex.: *foreningens formand* (lit.: the association's president) REL_GEN
  formanden for foreningen (the president of the association) PREPOSITIONAL_OBJECT

Simple, abstract nouns

- Properties may denote one of the following four property categories (cf. Lenci et al. 2000:228-241): quality, social property, physical property and psychical property.
  Ex.: *husets farve* (the colour of the house)

Deverbal nouns
Nouns derived from verbs often subcategorise for a subjective or an objective genitive and may also subcategorise for a prepositional phrase.
Ex.: *Hansens erkendelse af, at der var en fejl*
(lit.: Hansen's acknowledge of that there was a mistake)

In all other cases, the genitive is not encoded as complement, especially because of the fact that it is very difficult to differentiate between relational and possessive genitives.
Ex.: *landets sprog* (lit.: the country's language)

2.2.6 Noun as a complement of mass entity nouns

Mass entity nouns may be of three different subtypes: container, partitive and numeral.
It holds for each subtype that the complement is obligatory and it must be indefinite and undetermined, and the complements of container nouns must be countable nouns in plural. Partitive nouns allow only for nouns in singular. Numerals allow only for countable nouns in plural.
Ex.: *en kasse æbler* (a box of apples) : container nouns
et stykke kage (a piece of cake): partitive noun
et tusinde år (one hundred years): numeral
Their syntactic construction is described in the database in the following way:
(Self: definite_suffix_allowed: NO,
Position: 1; Function: SPEC_N; Optional: NO; Syntactic_Label: NP; Casus: UNMARKED;
NpIndex: N)

2.2.7 Prepositional phrases

Nouns, like verbs may subcategorise for one or more prepositional phrases (the governed elements may be: NP, that-clause, interrogative clause, wh-clause or infinitive) with the function PREPOSITIONAL_OBJECT. Accordingly, the method of encoding of prepositional complements in noun valency patterns is similar to that of verbs, below a few illustrating examples are provided.

- Mass nouns
  Ex.: *flertallet af indbyggerne* (the majority of the inhabitants)
- Semiotic artifacts
  bogen om landbruget (the book about agriculture)
- Direction nouns
  Ex.: *stien fra huset til stranden* (the path from the house to the beach)

2.2.8 Clausal complements

A number of nouns subcategorises for clausal complements: a that-clause or an infinitive.
Ex.: *Den metode at anvende datamater til forskning*
(lit.: the method to use computers for research): INFINITIVE
Den kendsgerning at han aldrig kommer for sent
(lit.: the fact that he never comes too late): THAT_CLAUSE

In the example provided below, the noun *diskussion* illustrates the four different phrase types governed by the preposition '*om*' (about):
Ex.: *publikums diskussion med panelet om den nye lov*: NP
(lit.: the audience's discussion about the new act)
diskussionen om, at slutningen nok var ironisk: THAT_CLAUSE
(lit.: the discussion about that the end might be ironic)
diskussionen om, hvorfor færre syge danskere kan se frem til..: WH-CLAUSE
(lit.: the discussion about why fewer Danes can look forward to…)
præsidentens diskussioner med rådgivere om at bryde ind: INFINITIVE
(lit.: the president's discussion with advisers about to intervene)

This is reflected by a complex description of these constructions
Position: 1; Function: REL_GEN; Optional: YES; Syntactic_Label: NP; Casus:

GENITIVE; Position: 2; Function: PREPOSITIONAL_OBJECT, Optional:YES, Syntactic_Label: PP, Introducer: MED, Clause_Type: NULL; NP_Type: N; Ccontrol: NOCONTROL; Coref: NOCOREF;
Position: 3, Function; PREPOSITIONAL_OBJECT; Optional:YES, Syntactic_label: PP; Introducer:OM; Clause_Type: NULL; NP_Type: N; Control: NOCONTROL; Coref: NOCOREF/
Introducer: OM; Clause_Type: THAT_CLAUSE; NP_Type: NULL; Control: NOCONTROL; Coref: NOCOREF/
Introducer: OM; Clause_Type: WH-CLAUSE, NP_Type: NULL; Control: NOCONTROL; Coref: NOCOREF/
Introducer: OM; Clause_Type: INFINITIVE; NP_Type: NULL; Control: WITHOUTCONTROL; Coref: WITHOUTCOREF)

2.2.9 Syntactic units and noun alternations

A word that has two valency patterns, is generally encoded in two syntactic units (SynU's), each with one syntactic description, e.g. in the case of different prepositions.

Ex.: Syntactic unit 1: *variation af temaet* (variation of the theme)
Syntactic unit 2: *variation over temaet* (variation on the theme)

In case of homographs, the different semantics of the same syntactic realisation does not imply two syntactic units. E.g. the word *krampe* is encoded in one syntactic unit though it has two senses 'staple' and 'convulsion' since the syntactic realisation of the two senses is identical.

The following noun alternation types are encoded in STO (In the examples below, the description identifiers 'Dn…' reflecting the alternation are shown for the sake of clarity.)

- Genitive alternation

Two syntactic descriptions can be encoded in one syntactic unit if the two constructions are alternations, i.e. the semantic reading of the constructions is the same though the syntactic expressions differ.

Ex.: *forpagter* (tenant)
Syntactic unit 1
Dn1Pn-af *forpagteren af gården* (the tenant of the farm)
Dn1G *gårdens forpagter* (lit: the farm's tenant)

- Reciprocal constructions

Ex.: *forlovelse* (engagement)
Syntactic unit 1
Dn1G *Peter og Susannes forlovelse* (Peter and Susanne's engagement)
Dn1Pn-mellem *forlovelsen mellem Peter og Susanne*
(the engagement between P. and S.)
Dn2GPn-med *Peters forlovelse med Susanne*
(Peter's engagement with S.)

- Family relations

Ex.: *bror* (brother)
Syntactic unit 1
Dn1G *Jens' bror* (Jens' brother)

Dn1P-til *En bror til Jens* (lit: a brother to Jens)

2.3 Adjectives

This section describes the syntax of adjectives. Firstly, a general introduction to the syntactic encoding of adjectives is given. Secondly, the distinction between complements and adjuncts/free modifiers is briefly discussed. Thirdly, a list of syntactic functions of the complements is provided and commented. Fourthly, the arity of adjectives is discussed and a specification of the particular elements of the valency patterns is given. Finally, the valency patterns of adjectives are illustrated by prototypical examples.

2.3.1 The syntactic encoding of adjectives

Normal adjectives (i.e. all adjectives except quantifiers and numerals) are provided with one or more syntactic descriptions. The syntactic description of an adjective contains syntactic information about the adjective itself and about its valency pattern(s).

Adjectives can be used in a given construction attributively or predicatively. This feature is encoded in the *Self* element. Thus, adjectives that are used both in attributive and predicative function are connected to two *Self* elements.

2.3.2 The valency of adjectives

Obligatory and optional complements are encoded, while free modifiers in accordance with the general strategy adopted are not registered. For identification of valency-bound complements, various linguistic tests were used (e.g. such as topicalisation tests). In such cases where the distinction based on those tests was not clear enough, the element was regarded as a 'middle' and its encoding followed the same guidelines as those described in the section on verbs.

Ex.: *de var glade for deres venner* (they were happy with their friends):
complement encoded
de var glade fra morgenstunden (they were happy from early morning):
modifier not encoded.

2.3.3 The arity of adjectives and numbering of the positions in constructions

The syntactic description of adjectives records the syntactic elements the adjective can combine with. However, we also provide the arity information giving the number of valency-bound semantic arguments occurring in a given construction. This feature may be relevant for encoders especially if they want to add semantic information. Arity is registered in the identifier of the *Description* and *Construction* elements, e.g. Da3... or Ca3… . The following adjective arity types are registered: mono, di- and trivalent (cf. examples below).
In the following, the **external** and **internal** argument types are exemplified and the strategy adopted for counting of arity in adjective valency patterns is explained and illustrated.

The **external** argument of an adjective is the noun which is modified by the adjective within an attributive construction is the argument of that adjective at the semantic layer. The same holds for the noun phrase of predicative constructions. Thus, the external argument noun is included into the number of arity being the first element of the valency frame. For this reason, the example below is

monovalent. In the valency frame it occupies the Position 1, and its syntactic function is EXTERNAL_COMP.
Ex.: *en ny bil* (a new car): attributive structure with external argument '*bil'*
. *bilen er ny* (the car is new): predicative structure with external argument *'bil'*

Internal arguments are complements that are subcategorised for by the adjective; the example below contains an external and an internal argument, thus it is divalent. The external argument (NP: Peter) in Position 1 is treated in the above described way. It occupies the Position 2 is and it has the function PREPOSITIONAL_OBJECT; the preposition '*for*'(with) governs an NP.
Ex.: *Peter er glad for sin nye bil* (Peter is happy with his new car)

2.3.4 Predicative constructions with clausal complements

A few adjectives may have clausal complements (syntactic function: CLAUSCOMP), and these complements may also occur in extraposed construction, this means that the sentential subject appears in final position, the subject position, Position 1 is filled by a FORMAL_COMP, which does not correspond to any semantic argument and is not counted in the arity value, thus the example below is monovalent.
Ex.: *at ryge er farligt* (lit.: to smoke is dangerous)
det er farligt at ryge (lit.: it is dangerous to smoke)

2.3.5 Optionality of complements

Two types of optionality are distinguished in STO:

- Complements being syntactically optional, but semantically obligatory (i.e. the argument is semantically implied)
  Ex.: *de forskellige meninger* (the different opinions),
  viz. it is semantically implied that there are at least two meanings that are different from each other in some respects)
  Vi er forskellige (fra hinanden) (we are different (from each other))
  Hans æbler er forskellige (fra de andre) (i smag)
  (His apples are different (from the others) (in taste))
- Complements being both syntactically and semantically optional
  Ex.: *pigen er sød (mod dyrene)/(ved børnene)/(over for sine venner)*
  (the girl is nice (to the animals/children/her friends) viz. the complementation is optional as the adjective does not imply a semantic argument.

In the valency frame, an optional complement may have the value YES or YES_GEN for the feature OPTIONAL. The value YES_GEN is applied if the optional complement gives rise to more than one single description of the same construction as explained below.

2.3.6 Optionality in syntactic units and descriptions

The following examples specify the encoding of the various combinations of complements and optionality in appropriate syntactic units.

(a) Adjectives with complements being syntactically optional, but semantically obligatory are provided with one single divalent valency frame comprising the optional complement. Usually, this type of adjectives subcategorizes for one particular preposition (e.g. *'for'*), thus the adjective is

described with one single syntactic unit which reflects the construction with an optional PREPOSITIONAL_OBJECT in Position 2, the value of the Optional element is YES.
Ex.: *Børnene er bange (for hunden)* (The children are afraid (of the dog))

(b) Adjectives with complements being syntactically and semantically optional are provided with one monovalent valency frame without the optional complement, i.e. the adjective is described with one single syntactic unit which reflects this construction.
Ex.: *pigen er sød* (the girl is nice)

Further, the adjective is provided with one (or more) valency frame(s) comprising the optional complement(s). Often, this type of adjectives subcategorizes for more than one particular preposition (e.g. *'ved','mod'* and *'over for'* see example below). Accordingly, an appropriate number of syntactic units (comprising one description each) describe the various constructions. (The relationship between a syntactic unit and description(s) is presented in Chapter 3, The data). The adjective *'sød'* subcategorizes for a prepositional phrase which may be introduced by three different prepositions, thus the adjective is provided with three syntactic units describing one particular construction each. The optional complement is always a PREPOSITIONAL_OBJECT in Position 2, the value of the Optional element is YES_GEN.

Ex.: *pigen er sød (mod dyrene)* (the girl is nice (to/ towards the animals)
pigen er sød (ved børnene) (the girl is nice (to/ towards the children)
pigen er sød (over for sine venner) (the girl is nice (to/ towards her friends)

According to the general method adopted in STO, syntactic differences between constructions of a word must be reflected by different descriptions. This results in more syntactic units if the optional complement of an adjective can be realized with various syntactic structures.
In the example below, if the prepositional phrase (viz. the optional complement) is represented in the construction, there is a prepositional object control in evidence (viz. the NP contained in the prepositional phrase is the subject of the infinitive clause.) Of course, this is not the case if the optional complement is omitted.
Ex.: (a) *det er godt (for hende) at rejse* (lit.: It is good (for her) to travel)
(b) *at rejse er godt (for hende)* (lit.: to travel is good (for her))

In order to treat the control phenomenon appropriately, four descriptions are needed in STO: two for constructions with the prepositional complement incl. the appropriate control and different syntactic structures of the obligatory complement, and accordingly two without the optional complement.
Ex.: (a) *det er godt for hende at rejse* (with control)
(b) *det er godt at rejse* (without control)
(c) a*t rejse er godt for hende* (with control)
(d) a*t rejse er godt* (without control)

2.3.7 Syntactic functions of complements

The internal arguments of adjectives have various syntactic realizations, these are the following:

- ACOMP denotes nominal complements
- Ex.: *Han er hende taknemmelig* (lit.: He is her grateful)
- PREPOSITIONAL_OBJECT denotes prepositional complements
  Ex.: *Han er god til at tegne* (lit.: He is good to draw)

Maria er bange for hunde og heste (Maria is afraid of dogs and horses)

- CLAUSCOMP denotes clausal complements
  Ex.: *Det er godt at du er kommet/At du er kommet er godt*
  (lit.: It is good that you came/ That you came is good)
- FORMAL_COMP denotes formal subjects
  Ex.: *Det er usædvanligt at ryge* (lit.: It is unsual to smoke)
- EXTERN_COMP bruges for eksterne argumenter
  Peter er glad; de glade børn (Peter is happy; the happy children)

2.3.8 Valency patters of adjectives

In the following, the most representative valency frame types are presented. For many of these frames we also provide the most relevant features encoded in the syntactic description.

Monovalent adjectives

1. Valency frames with an external argument:

- Adjectives that can be used attributively:
  Ex.: *Annes røde bil* (Anne's red car)
  (Adj_Function: ATTRIBUTIVE;
  Position: 1; Function: EXTERN_COMP; Optional: NO; Syntactic_Label: NP; Casus: UNMARKED)

- Adjectives that can be used predicatively:
  Ex.: *Annes bil er rød* (Anne's car is red)
  (Adj_Function: PREDICATIVE;
  Position: 1; Function: EXTERN_COMP; Optional: NO; Syntactic_Label: NP; Casus NON_GENITIVE)

2. Valency frames without an external argument
   A Clause is the syntactic subject. Semantically this subject corresponds to an internal argument. These valency frames have an alternate frame representing the corresponding extraposed constructions. The formal subject of the extraposed constructions is the expletive *det.* The Clause in these constructions corresponds still to an internal argument in a semantic representation.

- Frames for constructions without extraposition:
  Ex.: *at han kom i dag, er tilfældigt* (lit.: that he came today is accidental)
  (Adj_Function: PREDICATIVE;
  Position: 1; Function: CLAUSECOMP; Optional:NO; Syntactic_Label: CLAUSE;
  Clause_Type:THAT_CLAUSE; Control: NOCONTROL; Coref: NOCOREF)
  Ex.: *[at du kom/at købe en smoking] er fint* (lit.: that you came/to buy a dinner jacket is fine)
  (Adj_Function: PREDICATIVE;
  Position: 1; Function: CLAUSECOMP; Optional:NO; Syntactic_Label: CLAUSE;
  Clause_Type: THAT_CLAUSE; Control: NOCONTROL; Coref: NOCOREF;
  Syntactic_Label: CLAUSE; Clause_Type:INFINITIVE; Control: WITHOUTCONTROL;
  Coref: WITHOUTCOREF)

- Frames for the corresponding extraposed constructions:
  Ex.: *det er tilfældigt at han kom i dag* (lit.: it is accidental, that he came today)

(Adj_Function: PREDICATIVE;
Position: 1; Function: FORMAL_COMP; Optional:NO; Syntactic_Label: NP: expletive: DET; Casus: NOMINATIVE;
Position: 2; Function: CLAUSECOMP; Optional:NO; Syntactic_Label: CLAUSE; Clause_Type:THAT_CLAUSE; Control: NOCONTROL; Coref: NOCOREF)

Divalent adjectives

1. Valency frames with an external nominal argument and an internal argument which is a prepositional phrase. This prepositional phrase may be obligatory or optional:
   - The complement within the prepositional phrase is a noun (this is the default value)
     Ex.: *jeg er tom for ideer* (lit.: I'm empty of ideas)
     (Adj_Function: PREDICATIVE;
     Position: 1; Function: EXTERN_COMP; Optional:NO; Syntactic_Label: NP; Casus: NON_GENITIVE ;
     Position: 2; Function: PREPOSITIONAL_OBJECT; Optional:NO; Syntactic_Label:PP; Introducer: for; NP_Type: N)
     Ex.: *huset ligger bekvemt (for hendes arbejde)* lit.:(the house is situated favourably (for her work)
     (Adj_Function: PREDICATIVE;
     Position: 1; Function: EXTERN_COMP; Optional:NO; Syntactic_Label: NP; Casus: NON_GENITIVE;
     Position: 2; Function: PREPOSITIONAL_OBJECT; Optional:YES_GEN; Syntactic_Label:PP; NP_Type: N)

   - The complement within the prepositional phrase is a noun phrase or an infinitive:
     Ex.: *han er hurtig til [madlavning/at lave mad]*
     (lit.: he is quick to [cooking/to cook])
     (Adj_Function: PREDICATIVE;
     Position: 1; Function: EXTERN_COMP; Optional:NO; Syntactic_Label: NP; Casus: NON_GENITIVE; NPIndex: INDEXI;
     Position: 2; Function: PREPOSITIONAL_OBJECT; Optional:YES_GEN; Syntactic_Label:PP;
     Introducer: TIL; NP_Type: N; Clause_Type: NULL; Control:NOCONTROL; Coref: NOCOREF/
     Introducer: TIL; NP_Type: NULL; Clause_Type: INFINITIVE; Control: SUBJECTCONTROL; Coref: COI)

   - Further divalent types with an external nominal argument and an internal argument which is a prepositional phrase
     - The complement within the prepositional phrase is an infinitive with subject control
     - The complement within the prepositional phrase is a noun phrase, a that-clause or an infinitive with subject control.
     - The complement within the prepositional phrase is a noun phrase, a that-clause or an interrogative clause.

2. Valency frames with an external argument and an internal argument which is a Clause:
   Ex.: *Engelsk er let at lære* (lit.: English is easy to learn)
   (Adj_Function: PREDICATIVE;

Position: 1; Function: EXTERN_COMP; Optional:NO; Syntactic_Label: NP; Casus: NON_GENITIVE; NPIndex: INDEXI;
Position: 2; Function: CLAUSECOMP; Optional:NO; Syntactic_Label:CLAUSE; Clause_Type: INFINITIVE; Control: WITHOUTCONTROL; Coref: WITHOUTCOREF)

3. Valency frames without an external argument. The syntactic subject is a clause and the adjective subcategorizes also for a prepositional phrase. These adjectives may also occur in extraposed constructions.

A construction without extraposition:
Ex.: *at redde barnet var modigt af ham*
(lit.: to save the child was courageous of him)
(Adj_Function: PREDICATIVE;
Position: 1; Function: CLAUSECOMP; Optional:NO; Syntactic_Label: CLAUSE; Clause_Type: INFINITIVE; Control: PREPOBJECTCONTROL; Coref: CON; Position: 2; Function: PREPOSITIONAL_OBJECT; Optional: NO; Syntactic_Label:PP; Introducer: af; NPIndex: N; Clause_Type: NULL; Control: NOCONTROL; Coref: NOCOREF)

The corresponding extraposed construction:
Ex.: *det var modigt af ham at redde barnet*
(lit.: it was courageous of him to save the child)
(Adj_Function: PREDICATIVE;
Position: 1; Function: FORMAL_COMP; Optional:NO; Syntactic_Label: NP: expletive: DET; Casus: NOMINATIVE;
Position: 2; Function: PREPOSITIONAL_OBJECT; Optional: NO; Syntactic_Label:PP;
Introducer: af; NPIndex: N; Clause_Type: NULL; Control:NOCONTROL; Coref: NOCOREF;
Position: 3; Function: CLAUSECOMP; Optional:NO; Syntactic_Label: CLAUSE; Clause_Type: INFINITIVE; Control: PREPOBJECTCONTROL; Coref: CON)

Trivalent adjectives

1. Adjectives with two prepositional phrases, in the second prepositional phrase the complement may be a noun phrase, a that-Clause, an interrogative Clause or an infinitive.
Ex.: *han var enig (med moderen) (om [flytningen/at de skulle flytte/ hvornår de skulle flytte/at flytte])*
(lit.: he was agreeing (with the mother) on ([the removal/that they should move/when they should move/to move]. (NB: The adjective '*enig*' is translated into English by the verb 'agree'.)
(Adj_Function: PREDICATIVE;
Position: 1; Function: EXTERN_COMP; Optional:NO; Syntactic_Label: NP; Casus: NON_GENITIVE; NPIndex: INDEXI;
Position: 2; Function: PREPOSITIONAL_OBJECT; Optional: YES; Syntactic_Label:PP; Introducer: MED; NP_Type: N; Clause_Type: NULL; Control: NOCONTROL; Coref: NOCOREF;
Position: 3; Function: PREPOSITIONAL_OBJECT; Optional: YES_GEN; Syntactic_Label:PP;
Introducer: OM; NP_Type: N; Clause_Type: NULL; Control: NOCONTROL; Coref: NOCOREF;

Introducer: OM; NP_Type: NULL; Clause_Type: THAT_CLAUSE; Control: NOCONTROL; Coref: NOCOREF;
Introducer: OM; NP_Type: NULL; Clause_Type: INTERROGATIVE_CLAUSE; Control: NOCONTROL; Coref: NOCOREF;/
Introducer: OM; NP_Type: NULL; Clause_Type: INFINITIVE; Control: SUBJECTCONTROL; Coref: COI)

2. Adjectives with one nominal complement and one prepositional phrase. The prepositional complement may be a noun phrase, a that-Clause or an infinitive.
Ex.: *De var (mig) behjælpelig (med [kufferterne /at kufferterne blev hentet/at hente kufferterne])*
(lit.: They were (me) assistant (with [the luggage/ that the luggage was brought/to bring the luggage]))

2.3.9 Syntactic units and alternations

As it has been briefly mentioned in the introductory section on basic principles and description elements, syntactic alternations are encoded in the same syntactic unit. For adjectives, the set of alternations encoded as such is limited to cases where the adjectives occur in constructions, where the syntactic subject is a clause. These adjectives may also occur in corresponding extraposed constructions, where the syntactic subject is the expletive *det*, e.g.
Alternation within one Syntactic unit:
Ex.: D1t - *At han kom i dag, er tilfældigt* (lit. That he came today is fortuitous)
Da1ext - *Det er tilfældigt at han kom i dag* (lit. It is fortuitous that he came today)

An exception is the following alternation type, where only the first construction is encoded in STO:
Ex.: *Skibet er let at styre*/*Et let skib at styre*
(lit. The ship is easy to steer/an easy ship to steer): two alternate constructions
Da2Pi – *Skibet er let at styre*: the construction encoded

Constructions such as those in the following example are not coded as alternations, the two syntactic descriptions occur in two different syntactic units:
Ex.: *Hans mening er sagen uvedkommende/Hans mening er uvedkommende for sagen*
(lit.: His opinion is the subject irrelevant/His opinion has not bearing on the subject)
Syntactic unit 1: Da1exN - *Hans mening er sagen uvedkommende*
Syntactic unit 2: Da1exP-for - *Hans mening er uvedkommende for sagen*.

3 The data: STO Syntax represented as XML elements

This chapter describes the structure of the XML files in which the STO syntax is encoded (Section 1), then the number of relevant XML elements is provided (Section 2) and, finally, an example of syntactic encoding in XML is given (Section 3).

3.1 The Structure of the STO Syntax XML Files

The root element in the lexicon's XML file is *STO_Syntax*. The root element contains the following five elements:

1. Morph_Syn_Units
2. Descriptions
3. Selfs
4. Constructions
5. Phrases

- The *Morph_Syn_Units* element contains a sequence of *Mu_Synu* elements which connect STO syntactic units to STO morphological units through their identifier.
- The *Descriptions* element contains a sequence of *Description* elements.
- The *Selfs* element contains a sequence of *Self* elements.
- The *Constructions* element contains a sequence of *Construction* elements.
- The *Phrases* element contains a list of phrases (*NP, PP, Clause, ADVP* and *AP* elements).

In the following we describe how syntax is represented in STO through these XML elements.

3.1.1 Mu_Synu, Mu_Id and Spelling elements

Mu_Synu elements connect the syntactic descriptions of words, expressed in *Synu* elements with the identifiers of the morphological descriptions of the same words (*Mu_id* elements).
A *Synu* element stands for a syntactic unit and contains a description of a syntactic pattern of a word. A word may occur in different syntactic contexts, thus the same morphological unit can be bound to more *Synu* elements. The *Mu_Synu* element also contains one or more *Spelling* elements which give the spelling(s) of the described word.
The structure of the *Mu_Synu* elements is given in Figure 1.

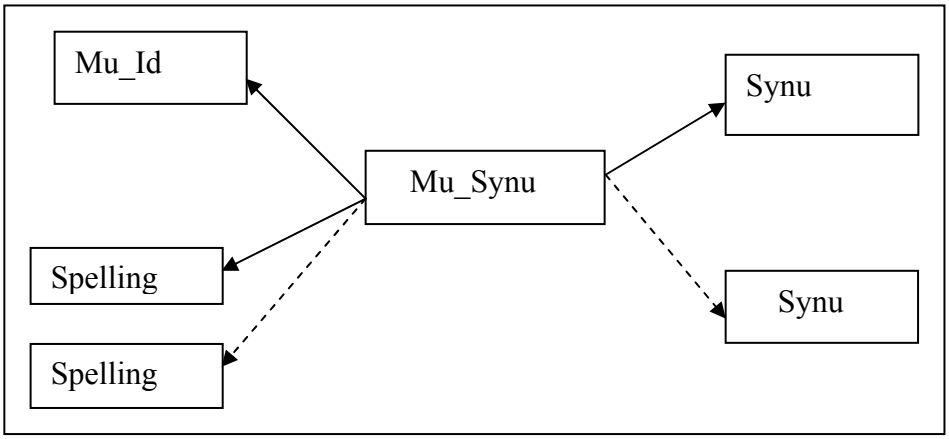


Figure 1: The structure of Mu_Synu elements

An example of an Mu_Synu is given in (1).

(1)

```
<Mu_Synu>
            <Mu_Id>"AFBETALE"</Mu_Id>
                        <Spelling>afbetale</Spelling>
            <Synu Id="SYNU_AFBETALE_1">
            …
            </Synu>
            <Synu Id="SYNU_AFBETALE_2">
                        ……
            </Synu>
```

```
</Mu_Synu>
```

In (1) the word with Spelling “afbetale” and with the Mu_Id “AFBETALE” has two syntactic readings encoded in the two Synus with the identifiers “SYNU_AFBETALE_1” and “SYNU_AFBETALE_2”.

3.1.2 Synu, Description and Construction elements

A syntactic pattern may have different realizations, thus a syntactic unit can contain more syntactic descriptions (*Description* elements). The binding between a syntactic unit and a description is done through the *Synu_Description* elements. A syntactic description consists of the characteristics of the actual word (the head) in the syntactic pattern (*Self* element) and the valency pattern (*Construction* element). The structure of a syntactic unit (*Synu*) is presented in figure 2.

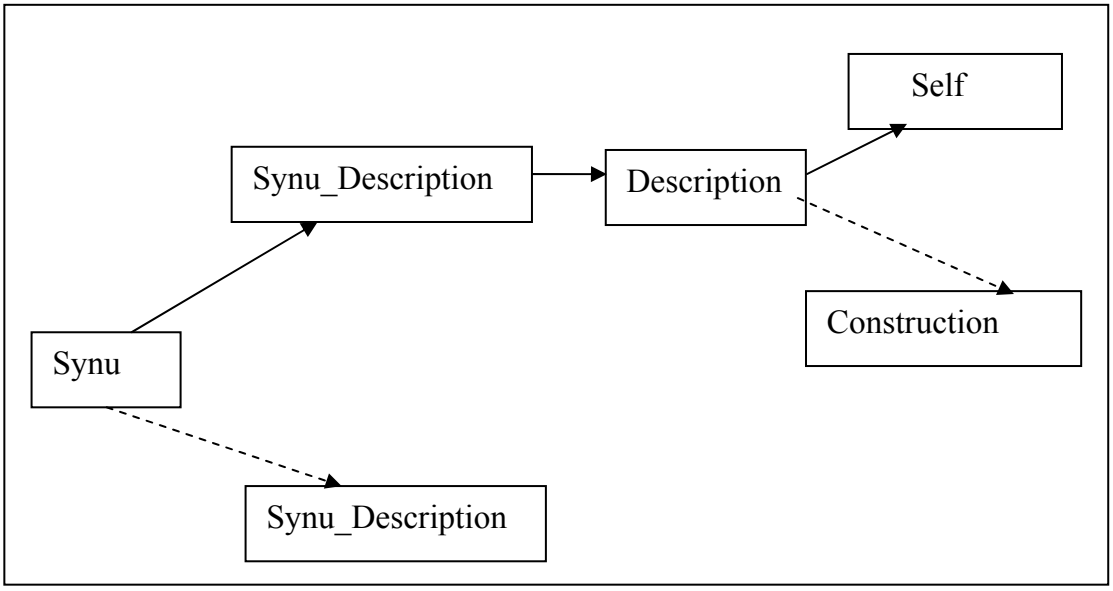


Figure 2: The structure of a syntactic unit (Synu)

An example of these elements in XML is given in (2).

(2)

```
<Synu Id="SYNU_AFBETALE_1">
           <Synu_Description Description_Id="Dv2N">
                      <Example> fordi vi først skal afbetale gælden.</Example>
           </Synu_Description>
</Synu>
<Synu Id="SYNU_AFBETALE_2">
           <Synu_Description Description_Id="Dv2P-på">
                      <Example> i stand til selv at købe og afbetale på
           huset</Example>
           </Synu_Description>
</Synu>

……

<Description Id="Dv2N" Construction_Id="Cv2N" Self_Id="have_NO_NO">
           <Naming>divalent:  NP, obligatory, NP</Naming>
```

```
            <Example>   ..når de beregner den tre-delte statsskat med
            skattesatser på 6, 12, og ca. 50% Korpuskilde:
            BGH\AVIS40.88</Example>
</Description>
```

In (2), the two syntactic units from example (1) are shown, together with the *Description* for the first unit. The first syntactic unit refers to a description "Dv2N" which, as explained in the *Naming* element, describes a divalent construction (the word subcategorizes for two constituents). The *Example* element under the *Description* contains a corpus example of a divalent syntactic pattern. The description refers to a Self element via the identifier "have_NO_NO" and a *Construction* element via the identifier "Cv2N". The *Self* element says that the word described is a verb and that it takes "have" (have) as auxiliary. The *Example* element may contain either a standard example illustrating the construction type but not necessarily the particular entry word, or an individual example containing the entry word itself in its narrow context.

A valency pattern (*Construction* element) is described by the canonical position of its constituents (*Position* element), the syntactic function (*Function* elements) of this position, the optionality status of the position (*Optional* element) and the syntactic characteristics of the phrases which may occur in that position (*Constituent* elements). The constituents may be phrases of different types: nominal phrases (*NP* elements), clauses (*Clause* elements), prepositional phrases (*PP* elements), adjectival phrases (*AP* elements) and adverbial phrases (*ADVP* elements). The structure of a *Construction* is given in figure 3.

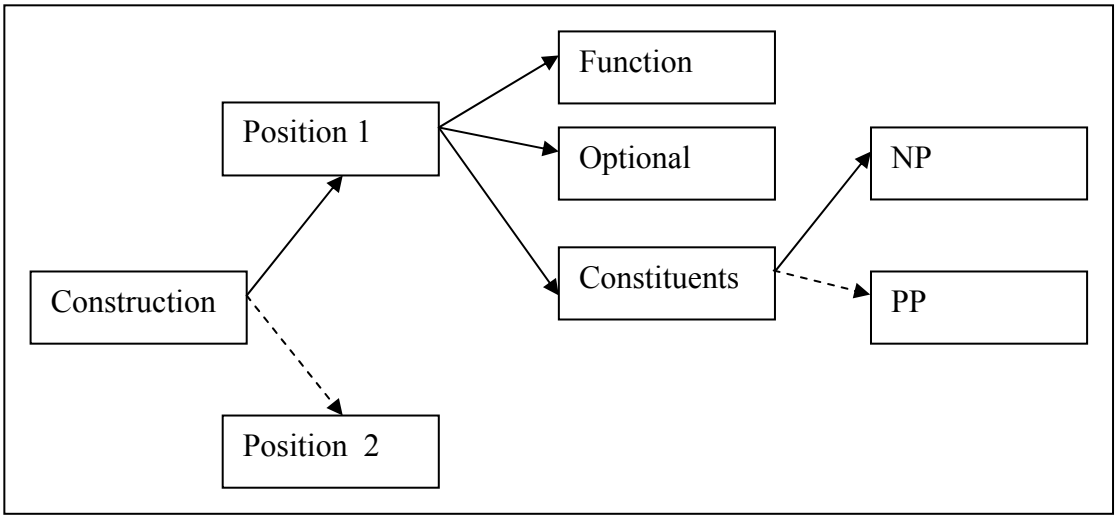


Figure 3: The structure of a Construction.

An XML example containing the Construction and the phrases from (2) is presented in (3) below.

(3)

```
<Construction Id="Cv2N">
            <Position Number="1">
                        <Function>SUBJECT</Function>
                        <Optional>NO</Optional>
                        <Constituents>
                                    <Constituent Syntactic_Label="NP"
Phrase_Id="NOM"/>
                        </Constituents>
```

```
            </Position>
            <Position Number="2">
                        <Function>OBJECT</Function>
                        <Optional>NO</Optional>
                        <Constituents>
                                    <Constituent Syntactic_Label="NP"
Phrase_Id="ACC"/>
                        </Constituents>
            </Position>
</Construction>
```

The *Construction* "Cv2N" in (3) consists of two positions: the first position has the syntactic function of subject, is obligatory (the *Optional* element contains the value "NO") and refers to a phrase which is a nominal phrase. The *Phrase* element contains the characteristics of nominal phrases that may occur in this position (e.g. the case of the nominal phrase must be NOMINATIVE). The second position is also obligatory, it has the syntactic function "OBJECT" and the phrase in it must also be a nominal (and its case must be accusative).

3.1.3 Self elements

There are different types of elements that describe the Self for verbs, adjectives, nouns and adverbs.

3.1.4 Self for verbs

The *Self* for verbs (*Self* element with *Cat* attribute having value VERB) contains the following elements: *Reflexive* (values YES, NO), *Particle* (NO, or particle if the verb is phrasal), *Auxiliary* (values HAVE or VÆRE), Passive (values YES, NO, or UNMARKED), *Modal* (values YES or NO), *Aux* (saying whether the verb is an auxiliary with the values YES or NO).

3.1.5 Self for nouns

The *Self* for nouns (*Self* element with *Cat* attribute having value NOUN) contains one element, *Definite_Suffix_Allowed*, which can take the values YES or NO.

3.1.6 Self for adjectives

The *Self* for adjectives (*Self* element with *Cat* attribute having value ADJECTIVE) contains one element *Adj_Func*, which can take the values PREDICATIVE or ATTRIBUTIVE.

3.1.7 Self for adverbs

The *Self* for adverbs (*Self* element with *Cat* attribute having value ADVERB) contains the following elements: *Modifying, Modifiable, Context, Appear_In_Predicative, Fundament_Field, Prefinit_Field, Nexus_Field, Manner_Field, Predicative_Field, Bound_Adverbial_Field, Prepositional_Object_Field, Final_Field, Pre_Or_Postmodifies, Pre_Or_Postmod_Inf, Appear_In_Question.*

The *Modifying* element describes the expressions that the adverb can modify and can contain the following string values:

- ADJECTIVE
- ADJECTIVE PHRASE
- ADVERB
- PREPOSITIONAL PHRASE
- IKKE (viz. negation expressed by 'no')

- NOUN PHRASE
- VERB
- VERB PHRASE
- SENTENCE
- EXC (viz. the formal subject)

The *Modifiable* element describes whether the adverb can be modified and whether it can be modified by the adverb *meget* (very) and it contains one of the following string values: YES, NO,YES_MEGET.

The *Content* element describes whether the adverb occurs in particular contexts and it can contain the following string values: NO_RESTRICTIONS, POSITIVE or NEGATIVE.

The *Appear_In_Predicative* element describes whether the adverb can occur in predicative constructions and it contains the string values YES or NO.

The *Fundament_Field* element describes whether the adverb can occur in the fundament field and it contains the string values YES, NO or NA (Not Applicable).

The *Prefinit_Field* element describes whether the adverb can precede a finite verb and it contains the string values YES, NO or NA (Not Applicable).

The *Nexus_Field* element describes where in the nexus the adverb can occur and can contain the following string values:

- ONLY_THEME
- ONLY_FOCUS
- BOTH
- NEITHER
- NA

The *Manner_Field, Predicative_Field, Bound_Adverbial_Field, Prepositional_Object_Field, Final_Field* elements describe whether the adverb can occur in the manner, predicative, prepositional object and final field, respectively. All of them can contain the string values YES, NO or NA (Not Applicable).

The *Pre_Or_Postmodifies* element indicates whether the adverb precedes or follows the modified expressions and it can take the following string values:

- PREMODIFIES
- POSTMODIFIES
- BOTH
- NA

The *Pre_Or_Postmod_Inf* element indicates whether the adverb precedes or follows the modified infinitive and it takes the same string values as the *Pre_Or_Postmodifies* element.

The *Appear_In_Question* element indicates whether the adverb can occur in questions and it can take the following string values:

- YES_ONLY
- NO

- NO_RESTRICTIONS
- NA

3.1.8 The elements describing valency patterns

The *Position* in a Construction can range from 1 to 4 (monovalent, divalent, trivalent and tetravalent constructions). Zerovalent constructions do not contain any *Position* element.

The *Function* element can contain one of the following string values:

- SUBJECT
- OBJECT
- INDIRECT_OBJECT
- PREPOSITIONAL_OBJECT
- OBJECT_PREDICATE
- FORMAL_SUBJECT
- ADVERBIAL
- SUBJECT_PREDICATE
- REL_GEN
- EXTERN_COMP
- FORMAL_COMP
- CLAUSCOMP
- ACOMP
- SPEC_N
- SOM_PP.

The *Optional* element contains one of the three string values: YES, NO, and YES_OVERG. YES_OVERG is used for the complements of adjectives and indicates that the description of the optional element can result in syntactic ambiguity.

In the following we describe the elements and the values which apply for each phrasal type.

NP elements

NP elements describe nominal phrases, and they are bound to elements describing the phrase's case, reflexivity, type (viz. expletive or no), definiteness and to an index (NPIndex) which is used to mark co-reference in constructions with control phenomena. (such as NPIndex=I coref=COI). The possible content of the elements mentioned above is given in the tables below.

ELEMENT	VALUES					
Case	NOMINATIVE	GENITIVE	ACCUSATIVE	ACCUSATIVE_INDIREKTE	NO_GENITIVE	UNMARKED

ELEMENT	VALUES		
Reflexive	YES	NO	UNMARKED

ELEMENT	VALUES		
Expletive	DET	DER	NO

ELEMENT	VALUES					
NPIndex	I	J	K	L	M	N

CLAUSE elements

Clause elements contain information about the clausal type, control and co-reference, expressed in the subelements *Clause_Type*, *Control* and *Coref.* The possible content of these elements is given in the following tables.

ELEMENT	VALUES			
Clause_Type	THAT_CLAUSE	INTERROGATIVE_CLAUSE	INFINITIVE	INFINITIVE_WITHOUT_INTRO

ELEMENT	VALUES			
Control	NOCONTROL	WITHOUTCONTROL	SUBJECTCONTROL	OBJECTCONTROL
		INDIRECTOBJECTCONTROL	PREPOBJECTCONTROL	RAISING

ELEMENT	VALUES							
Coref	NOCOREF	WITHOUTCOREF	COI	COJ	COK	COL	COM	CON

The values in the elements *Control* and *Coref* depend on the type of control. Only infinitive clauses can have control (value is not NOCONTROL).

PP

PP elements describe prepositional phrases through the subelements *Introducer, Clause_Type, NP_Type, Control, Coref*. The *Introducer* element has as its value the preposition which introduces the phrase. *Clause_Type* is reserved for prepositional phrases, where the complement is a clause. In this case the element *NP_Type* has the value NA (Not Applicable). If the prepositional complement is an NP, the *Clause_Type* element has value NA and *NPIndex* contain the index of the NP (as for the NP element). The values for the *Control* and *Coref* elements are the same as for clauses.

NOTE: *ADVP* and *AP* are empty.

3.2 The number of main XML elements

The number of the main elements in the STO Syntax lexicon are the following:

- Mu_Synu: 45309
- Synu: 57887
- Synu_Description: 60381
- Description: 1363
- Self: 103
- Construction: 985
- NP:11
- PP:113
- Clause: 9

3.3 An example of XML

In the following a complete example of the encoding of the verb afbetale is given as an example:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<STO_Syntax xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:noNamespaceSchemaLocation="STO_Syntax.xsd">
   <Morph_Syn_Units>
            …
            <Mu_Synu>
                        <Mu_Id>"AFBETALE"</Mu_Id>
                        <Spelling>afbetale</Spelling>
                        <Synu Id="SYNU_AFBETALE_1">
                           <Synu_Description Description_Id="Dv2N">
                        <Example> fordi vi først skal afbetale gælden.</Example>
                           </Synu_Description>
                        </Synu>
                        <Synu Id="SYNU_AFBETALE_2">
                           <Synu_Description Description_Id="Dv2P-på">
                              <Example> i stand til selv at købe og afbetale på huset
                                                </Example>
                     </Synu_Description>
                        </Synu>
            </Mu_Synu>
...
   </Morph_Syn_Units>
   <Descriptions>
...
            <Description Id="Dv2N" Construction_Id="Cv2N"   Self_Id="have_NO_NO">
               <Naming>divalent:  NP, obligatory, NP</Naming>
               <Example>   ..når de beregner den tre-delte statsskat med skattesatser
                        på 6, 12, og ca. 50% Korpuskilde:       BGH\AVIS40.88</Example>
            </Description>
            <Description Id="Dv2P-på"           Construction_Id="Cv2P-på"
            Self_Id="have_NO_NO">
               <Naming>divalent: NP, oblique PP prep=på</Naming>
               <Example>Det er som et kludetæppe, som jeg har nørklet på hele
            livet,</Example>
            </Description>
...
   </Descriptions>
   <Selfs>
...
            <Self Cat="VERB" Id="have_NO_NO">
                                    <Reflexive>NO</Reflexive><Particle>NO</Particle>
               <Auxiliary>have</Auxiliary><Passive>NO</Passive>
               <Modal>NO</Modal><Aux>NO</Aux>
            </Self>
...
   </Selfs>
   <Constructions>
...
            <Construction Id="Cv2N">
                        <Position Number="1">
                           <Function>SUBJECT</Function> <Optional>NO</Optional>
                           <Constituents>
                                    <Constituent Syntactic_Label="NP" Phrase_Id="NOM"/>
                           </Constituents>
                        </Position>
                        <Position Number="2">
                                    <Function>OBJECT</Function
            <Optional>NO</Optional>
                                    <Constituents>
                                       <Constituent Syntactic_Label="NP"
Phrase_Id="ACC"/>
                                    </Constituents>
                        </Position>
            </Construction>
            <Construction Id="Cv2P-på">
                        <Position Number="1">
                           <Function>SUBJECT</Function><Optional>NO</Optional>
                           <Constituents>
                                    <Constituent Syntactic_Label="NP" Phrase_Id="NOM"/>
                           </Constituents>
```

```
                        </Position>
                        <Position Number="2">

<Function>PREPOSITIONAL_OBJECT</Function><Optional>NO</Optional>
                          <Constituents>
                                    <Constituent Syntactic_Label="PP"
hrase_Id="PÅ_NP_NOC_NOC"/>
                                    </Constituents>
                        </Position>
            </Construction>
...
   </Constructions>
   <Phrases>
            <NP Id="NOM">
                        <Casus>NOMINATIVE</Casus><Reflexive>NO</Reflexive>
                        <Expletive>NO</Expletive><Definite>UNMARKED</Definite>
                        <NPIndex>I</NPIndex>
            </NP>
            <NP Id="ACC">
                        <Casus>ACCUSATIVE</Casus><Reflexive>UNMARKED</Reflexive>
                        <Expletive>NO</Expletive><Definite>UNMARKED</Definite>
                        <NPIndex>J</NPIndex>
            </NP>
            <PP Id="PÅ_NP_NOC_NOC">
                        <Introducer>på</Introducer><Clause_Type>NA</Clause_Type>
                                    <NP_Type>N</NP_Type><Control>NOCONTROL</Control>
                        <Coref>NOCOREF</Coref> </PP>
   </Phrases>
</STO_Syntax>
```

4 Appendix

4.1 The XML Schema file for the STO syntax, STO_Syntax.xsd:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
            <xs:annotation>
                        <xs:documentation>This is the export file of the syntax of the
STO                     Danish lexicon.</xs:documentation>
            </xs:annotation>
            <xs:element name="STO_Syntax">
                        <xs:annotation>
                                    <xs:documentation> The root element STO_Syntax
contains the                                                elements Morph_Syn_Units,
Descriptions, Selfs, Constructions                          and
Phrases</xs:documentation>
                        </xs:annotation>
                        <xs:complexType>
                                    <xs:sequence>
                                                <xs:element ref="Morph_Syn_Units" />
                                                <xs:element ref="Descriptions"/>
                                                <xs:element ref="Selfs"/>
                                                <xs:element ref="Constructions" />
                                                <xs:element ref="Phrases"/>
                                    </xs:sequence>
                        </xs:complexType>
            <!-- key and keyrefs giving restrictions on ids and their references  -->

                        <xs:key name="des_id">
                                    <xs:selector xpath="Descriptions/Description"/>
                                    <xs:field xpath="@Id"/>
                        </xs:key>
                        <xs:keyref name="des_id_ref" refer="des_id">
                                    <xs:selector

            xpath="Morph_Syn_Units/Mu_Synu/Synu/Synu_Description"/>
                                    <xs:field xpath="Description_Id"/>
                        </xs:keyref>
                        <xs:key name="cos_id">
                                    <xs:selector xpath="Constructions/Construction"/>
                                    <xs:field xpath="@Id"/>
                        </xs:key>
                        <xs:keyref name="cos_id_ref" refer="cos_id">
                                    <xs:selector xpath="Descriptions/Description"/>
                                    <xs:field xpath="Construction_Id"/>
                        </xs:keyref>
                        <xs:key name="s_id">
                                    <xs:selector xpath="Selfs/Self"/>
                                    <xs:field xpath="@Id"/>
                        </xs:key>
                        <xs:keyref name="s_id_ref" refer="s_id">
                                    <xs:selector xpath="Descriptions/Description"/>
                                    <xs:field xpath="Self_Id"/>
                        </xs:keyref>
                        <xs:key name="phr_id">
                                    <xs:selector xpath="Phrases/*"/>
                                    <xs:field xpath="@Id"/>
                        </xs:key>
                        <xs:keyref name="phr_id_ref" refer="phr_id">
                                    <xs:selector

            xpath="Constructions/Construction/Position/Constituents/Constituent"/>
                                    <xs:field xpath="Phrase_Id"/>
                        </xs:keyref>
            </xs:element>
            <xs:element name="Morph_Syn_Units">
                        <xs:annotation>
                                    <xs:documentation> The element Morph_Syn_Units
contains                                    Mu_Synu elements which connect STO
syntactic units to STO                                      morphological
units</xs:documentation>
                        </xs:annotation>
                        <xs:complexType>
```

```
                                    <xs:sequence>
                                                <xs:element maxOccurs="unbounded"
ref="Mu_Synu"/>
                                    </xs:sequence>
                        </xs:complexType>
            </xs:element>
            <xs:element name="Mu_Synu">
                        <xs:annotation>
                                    <xs:documentation> The element Mu_Synu connects the
id of a                             morphological unit (Mu_Id) and its
possible spellings                                  (Spelling element) to one or
more syntactic
                                    units (Synu). </xs:documentation>
                        </xs:annotation>
                        <xs:complexType>
                                    <xs:sequence>
                                                <xs:element name="Mu_Id"
type="Short_String_Type"/>
                                                <xs:element maxOccurs="30"
name="Spelling"
            type="Short_String_Type"/>
                                                <xs:element maxOccurs="unbounded"
ref="Synu"/>
                                    </xs:sequence>
                        </xs:complexType>
            </xs:element>
            <xs:element name="Synu">
                        <xs:annotation>
                                    <xs:documentation> The element Synu contains a
reference to                                        one or more syntactic
descriptions (Synu_Description).
                                    </xs:documentation>

                        </xs:annotation>
                        <xs:complexType>
                                    <xs:sequence>
                                                <xs:element maxOccurs="100"
name="Synu_Description"
            type="Synu_Description_Type"/>
                                    </xs:sequence>
                                    <xs:attribute name="Id" type="xs:ID" use="required"/>
                        </xs:complexType>
            </xs:element>
            <xs:element name="Example" type="Long_String_Type"/>
            <xs:element name="Descriptions">
                        <xs:annotation>
                                    <xs:documentation> The element Description contains
the                                 sequence of Description elements (i.e.
descriptions of                                     syntactic valency) contained
in the STO-lexicon.
                                    </xs:documentation>
                        </xs:annotation>
                        <xs:complexType>
                                    <xs:sequence>
                                                <xs:element maxOccurs="unbounded"
name="Description"
            type="Description_Type"/>
                                    </xs:sequence>
                        </xs:complexType>
            </xs:element>
            <xs:element name="Selfs">
                        <xs:annotation>
                                    <xs:documentation> The element Selfs  contains the
sequence                            of descriptions of valency heads (Self)
in the STO-base.
                                    </xs:documentation>
                        </xs:annotation>
                        <xs:complexType>
                                    <xs:sequence>
                                                <xs:element maxOccurs="unbounded"
name="Self"
            type="Self_Type"/>
                                    </xs:sequence>
                        </xs:complexType>
            </xs:element>
            <xs:element name="Constructions">
                        <xs:annotation>
```

```
                                    <xs:documentation> The element Constructions
contains the                                                    sequence of construction
elements (Construction) which                                                   express valency
descriptions without information about the
            valence head.</xs:documentation>
                        </xs:annotation>
                        <xs:complexType>
                                    <xs:sequence>
                                                <xs:element maxOccurs="unbounded"
name="Construction"                                             type="Construction_Type"/>
                                    </xs:sequence>
                        </xs:complexType>
            </xs:element>
            <xs:element name="Phrases">
                        <xs:annotation>
                                    <xs:documentation> The element Phrases contains the
sequence                            of Phrase elements constituting the syntactic
complements                                     (at the moment NP, PP, Clause).
                                     </xs:documentation>
                        </xs:annotation>
                        <xs:complexType>
                                    <xs:sequence>
                                                <xs:element minOccurs="0"
maxOccurs="unbounded"
            name="NP" type="NPP_Type"/>
                                                <xs:element minOccurs="0"
maxOccurs="unbounded"
            name="PP" type="PP_Type"/>
                                                <xs:element minOccurs="0"
maxOccurs="unbounded"
            name="Clause" type="ClauseP_Type"/>
                                    </xs:sequence>
                        </xs:complexType>
            </xs:element>
            <xs:simpleType name="Short_String_Type">
                        <xs:restriction base="xs:string">
                                    <xs:maxLength value="100"/>
                        </xs:restriction>
            </xs:simpleType>
            <xs:complexType name="Synu_Description_Type">
            <xs:annotation>
                                    <xs:documentation> An element of
Synu_Description_Type may
            contain an example (Example element) and a reference
                                    to the id of a Description.</xs:documentation>
                        </xs:annotation>
                        <xs:sequence>
                                    <xs:element minOccurs="0" ref="Example"/>
                        </xs:sequence>
                        <xs:attribute name="Description_Id" type="xs:IDREF"
                                                                        use="required"/>
            </xs:complexType>
            <xs:simpleType name="Long_String_Type">
                        <xs:restriction base="xs:string">
                                    <xs:maxLength value="1000"/>
                        </xs:restriction>
            </xs:simpleType>
            <xs:complexType name="Description_Type">
            <xs:annotation>
                        <xs:documentation> An element of type Description_Type may
contain                             an example (Example element) and may have a Name
(Naming).                                       It has a unique identifier (Id) and a
reference to a                                                          description of
the valency head (Self_Id) and to a valency
            pattern (Construction_Id)</xs:documentation>
                        </xs:annotation>
                        <xs:sequence>
                                    <xs:element minOccurs="0" name="Naming"

            type="Long_String_Type"/>
                                    <xs:element minOccurs="0" name="Example"
                                                                        type="Long_String_Type"/>
                        </xs:sequence>
                        <xs:attribute name="Id" type="xs:ID" use="required"/>
                        <xs:attribute name="Construction_Id" type="xs:IDREF"/>
                        <xs:attribute name="Self_Id" type="xs:IDREF" use="required"/>
            </xs:complexType>
```

```
            <xs:complexType name="Self_Type">
            <xs:annotation>
                        <xs:documentation> An element of type Self_Type has a unique
                                    identifier (Id) a category for the word and contains
a list of                           elements which vary if the valency head is a verb, a
noun, an                            adjective or an adverb.</xs:documentation>
                        </xs:annotation>
                        <xs:choice>
                        <!--  Self for Verbs -->
                                    <xs:sequence>
                                                <xs:element name="Reflexive"
type="YN_Type"/>
                                                <xs:element name="Particle"

            type="Short_String_Type"/>
                                                <xs:element name="Auxiliary"
type="Auxiliary_Type"/>
                                                <xs:element name="Passive"
type="YNU_Type"/>
                                                <xs:element name="Modal"
type="YN_Type"/>
                                                <xs:element name="Aux" type="YN_Type"/>
                                    </xs:sequence>
                                    <!--  Self for Substantives -->
                                    <xs:sequence>
                                                <xs:element
name="Definite_Suffix_Allowed"
                        type="YN_Type"/>
                                    </xs:sequence>
                                    <!--  Self for Adjectives -->
                                    <xs:sequence>
                                                <xs:element name="Adj_Func"
type="Adj_Func_Type"/>
                                    </xs:sequence>
                        <!--        Self for Adverbs -->
                                    <xs:sequence>
                                        <xs:element name="Modifying"
type="Modifying_Type" />
                                        <xs:element name="Modifiable"
type="Modifiable_Type" />
                                        <xs:element name="Context" type="Context_Type" />
                                        <xs:element name="Appear_In_Predicative"
                                                type="YN_Type" />
                                        <xs:element name="Fundament_Field"
type="YNNA_Type" />
                                        <xs:element name="Prefinit_Field"
type="YNNA_Type" />
                                        <xs:element name="Nexus_Field"
                                                type="Nexus_Field_Type" />
                                        <xs:element name="Manner_Field" type="YNNA_Type"
/>
                                        <xs:element name="Predicative_Field"
type="YNNA_Type" />
                                        <xs:element name="Bound_Adverbial_Field"
                                                                    type="YNNA_Type"
/>
                                        <xs:element name="Prepositional_Object_Field"
                                                                    type="YNNA_Type"
/>
                                        <xs:element name="Final_Field" type="YNNA_Type"
/>
                                        <xs:element name="Pre_Or_Postmodifies"

            type="Pre_Or_Postmodifies_Type" />
                                        <xs:element name="Pre_Or_Postmod_Inf"

            type="Pre_Or_Postmodifies_Type" />
                                        <xs:element name="Appear_In_Question"

            type="Appear_In_Question_Type" />
                                    </xs:sequence>
                        </xs:choice>
                        <xs:attribute name="Cat" type="Cat_Type" use="required"/>
                        <xs:attribute name="Id" type="xs:ID" use="required"/>
            </xs:complexType>
            <xs:simpleType name="Cat_Type">
                        <xs:restriction base="xs:string">
```

```
                                    <xs:enumeration value="ADJECTIVE"/>
                                    <xs:enumeration value="ADVERB"/>
                                    <xs:enumeration value="NOUN"/>
                                    <xs:enumeration value="VERB"/>
                        </xs:restriction>
            </xs:simpleType>
            <xs:simpleType name="YN_Type">
                        <xs:restriction base="xs:string">
                                    <xs:enumeration value="YES"/>
                                    <xs:enumeration value="NO"/>
                        </xs:restriction>
            </xs:simpleType>
            <xs:simpleType name="YNU_Type">
                        <xs:restriction base="xs:string">
                                    <xs:enumeration value="YES"/>
                                    <xs:enumeration value="NO"/>
                                    <xs:enumeration value="UNMARKED"/>
                        </xs:restriction>
            </xs:simpleType>
            <xs:simpleType name="YNNA_Type">
                        <xs:restriction base="xs:string">
                                    <xs:enumeration value="YES"/>
                                    <xs:enumeration value="NO"/>
                                    <xs:enumeration value="NA"/>
                        </xs:restriction>
            </xs:simpleType>
            <xs:simpleType name="Optional_Type">
            <xs:annotation>
               <xs:documentation> Type describing whether a complement is
                                    optional (YES or YES_OVERG), obligatory (NO).
YES_OVERG is only                                   used for adjectival complements and
indicates that removing the                                   optional complement results
in syntactic ambiguities, because the                                   given pattern is
already in the STO base.</xs:documentation>
                        </xs:annotation>
                        <xs:restriction base="xs:string">
                                    <xs:enumeration value="YES"/>
                                    <xs:enumeration value="NO"/>
                                    <xs:enumeration value="YES_OVERG"/>
                        </xs:restriction>
            </xs:simpleType>
            <xs:simpleType name="Auxiliary_Type">
            <xs:annotation>
              <xs:documentation> Type of auxiliary that the verb takes (have or
                        være?)</xs:documentation>
                        </xs:annotation>
                        <xs:restriction base="xs:string">
                                    <xs:enumeration value="have"/>
                                    <xs:enumeration value="være"/>
                        </xs:restriction>
            </xs:simpleType>
            <xs:simpleType name="Adj_Func_Type">
            <xs:annotation>
              <xs:documentation> The adjective appears in predicative or attributive
                                    position in the given Description.</xs:documentation>
                        </xs:annotation>
                        <xs:restriction base="xs:string">
                                    <xs:enumeration value="PREDICATIVE"/>
                                    <xs:enumeration value="ATTRIBUTIVE"/>
                        </xs:restriction>
            </xs:simpleType>

            <!-- Types describing the Self of adverbs -->
            <xs:simpleType name="Modifying_Type">
            <xs:annotation>
              <xs:documentation> Constituents that the adverb can modify
                                    </xs:documentation>
                        </xs:annotation>
                        <xs:restriction base="xs:string">
                                    <xs:enumeration value="ADJECTIVE"/>
                                    <xs:enumeration value="ADJECTIVE PHRASE"/>
                                    <xs:enumeration value="ADVERB"/>
                                    <xs:enumeration value="PREPOSITIONAL PHRASE"/>
                                    <xs:enumeration value="IKKE"/>
                                    <xs:enumeration value="NOUN PHRASE"/>
                                    <xs:enumeration value="VERB"/>
                                    <xs:enumeration value="VERB PHRASE"/>
```

```
                                    <xs:enumeration value="SENTENCE"/>
                                    <xs:enumeration value="EXC"/>
                        </xs:restriction>
            </xs:simpleType>
            <xs:simpleType name="Modifiable_Type">
            <xs:annotation>
               <xs:documentation> Can the adverb be modified?</xs:documentation>
                        </xs:annotation>
                        <xs:restriction base="xs:string">
                                    <xs:enumeration value="YES"/>
                                    <xs:enumeration value="NO"/>
                                    <xs:enumeration value="YES_MEGET"/>
                        </xs:restriction>
            </xs:simpleType>
            <xs:simpleType name="Context_Type">
            <xs:annotation>
                        <xs:documentation> Type of context in which the adverb must occur
    </xs:documentation>
                        </xs:annotation>
                        <xs:restriction base="xs:string">
                                    <xs:enumeration value="NO_RESTRICTIONS"/>
                                    <xs:enumeration value="POSITIVE"/>
                                    <xs:enumeration value="NEGATIVE"/>
                        </xs:restriction>
            </xs:simpleType>
            <xs:simpleType name="Nexus_Field_Type">
            <xs:annotation>
                        <xs:documentation> Where in the nexus field the adverb can occur
                        </xs:documentation>
            </xs:annotation>
                        <xs:restriction base="xs:string">
                                    <xs:enumeration value="ONLY_THEME"/>
                                    <xs:enumeration value="ONLY_FOCUS"/>
                                    <xs:enumeration value="BOTH"/>
                                    <xs:enumeration value="NEITHER"/>
                                    <xs:enumeration value="NA"/>
                        </xs:restriction>
            </xs:simpleType>
            <xs:simpleType name="Pre_Or_Postmodifies_Type">
                        <xs:restriction base="xs:string">
                                    <xs:enumeration value="PREMODIFIES"/>
                                    <xs:enumeration value="POSTMODIFIES"/>
                                    <xs:enumeration value="BOTH"/>
                                    <xs:enumeration value="NA"/>
                        </xs:restriction>
            </xs:simpleType>
            <xs:simpleType name="Appear_In_Question_Type">
                        <xs:restriction base="xs:string">
                                    <xs:enumeration value="YES_ONLY"/>
                                    <xs:enumeration value="NO"/>
                                    <xs:enumeration value="NO_RESTRICTIONS"/>
                                    <xs:enumeration value="NA"/>
                        </xs:restriction>
            </xs:simpleType>
            <xs:complexType name="Construction_Type">
            <xs:annotation>
               <xs:documentation> An element of Construction_Type has an unique
                        identifier referred to in the Description element
                        and describes valency elements by the contained Position
elements.
                         </xs:documentation>
            </xs:annotation>
                        <xs:sequence>
                                    <xs:element name="Position" type="Position_Type"
                                                            minOccurs="0"
maxOccurs="4"/></xs:sequence>
                        <xs:attribute name="Id" type="xs:ID" use="required"/>
            </xs:complexType>
            <xs:complexType name="Position_Type">
            <xs:annotation>
              <xs:documentation> An element of Position_Type gives the canonical
                        position of valency constituents (attribute Number) contained in
                                    the element Constituents, expresses the constituents
function                            (element Function) and indicates whether the
constituents are                                optional or obligatory in the element
Optional.               </xs:documentation>
            </xs:annotation>
```

```
                        <xs:sequence>
                                    <xs:element name="Function" type="Function_Type" />
                                    <xs:element name="Optional" type="Optional_Type" />
                                    <xs:element name="Constituents"
type="Constituents_Type" />
                        </xs:sequence>
                        <xs:attribute name="Number" type="PNumber_Type" use="required"/>
            </xs:complexType>
            <xs:simpleType name="PNumber_Type">
                        <xs:restriction base="xs:string">
                                    <xs:enumeration value="1"/>
                                    <xs:enumeration value="2"/>
                                    <xs:enumeration value="3"/>
                                    <xs:enumeration value="4"/>
                        </xs:restriction>
            </xs:simpleType>
            <xs:complexType name="Constituents_Type">
                        <xs:sequence>
                                    <xs:element name="Constituent"
type="Constituent_Type" maxOccurs="10"/>
                        </xs:sequence>
            </xs:complexType>
            <xs:complexType name="Constituent_Type">
            <xs:annotation>
              <xs:documentation> Syntactic label of a constituent and reference to
                          the constituent phrase for NP,PP and Clauses.
             </xs:documentation>
            </xs:annotation>
                        <xs:attribute name="Syntactic_Label" type="Syntactic_Label_Type"
                                    use="required"/>
                        <xs:attribute name="Phrase_Id" type="xs:IDREF"/>
            </xs:complexType>
            <xs:simpleType name="Syntactic_Label_Type">
                        <xs:restriction base="xs:string">
                                    <xs:enumeration value="NP"/>
                                    <xs:enumeration value="PP"/>
                                    <xs:enumeration value="CLAUSE"/>
                                    <xs:enumeration value="ADVP"/>
                                    <xs:enumeration value="AP"/>
                                    <xs:enumeration value="OTHER"/>
                        </xs:restriction>
            </xs:simpleType>
            <xs:simpleType name="Function_Type">
            <xs:annotation>
              <xs:documentation> List of the possible complement functions
            </xs:documentation>
            </xs:annotation>
                        <xs:restriction base="xs:string">
                                    <xs:enumeration value="ACOMP"/>
                                    <xs:enumeration value="ADVERBIAL"/>
                                    <xs:enumeration value="CLAUSCOMP"/>
                                    <xs:enumeration value="EXTERN_COMP"/>
                                    <xs:enumeration value="FORMAL_COMP"/>
                                    <xs:enumeration value="FORMAL_SUBJECT"/>
                                    <xs:enumeration value="INDIRECT_OBJECT"/>
                                    <xs:enumeration value="OBJECT"/>
                                    <xs:enumeration value="OBJECT_PREDICATE"/>
                                    <xs:enumeration value="PREPOSITIONAL_OBJECT"/>
                                    <xs:enumeration value="REL_GEN"/>
                                    <xs:enumeration value="SOM_PP"/>
                                    <xs:enumeration value="SPEC_N"/>
                                    <xs:enumeration value="SUBJECT"/>
                                    <xs:enumeration value="SUBJECT_PREDICATE"/>
                        </xs:restriction>
            </xs:simpleType>
            <xs:complexType name="NPP_Type">
            <xs:annotation>
              <xs:documentation> NP and its possible characteristics
                        </xs:documentation>
            </xs:annotation>
                        <xs:sequence>
                                    <xs:element name="Casus" type="Casus_Type" />
                                    <xs:element name="Reflexive" type="YNU_Type" />
                                    <xs:element name="Expletive" type="Expletive_Type" />
                                    <xs:element name="Definite" type="Definite_Type"/>
                                    <xs:element name="NPIndex" type="NPIndex_Type" />
```

```
                                    <xs:element name="Optionalfeature" type="xs:string"
                                                                        minOccurs="0"/>
                        </xs:sequence>
                        <xs:attribute name="Id" type="xs:ID" use="required"/>
            </xs:complexType>
            <xs:simpleType name="Casus_Type">
                        <xs:restriction base="xs:string">
                                    <xs:enumeration value="NOMINATIVE"/>
                                    <xs:enumeration value="ACCUSATIVE"/>
                                    <xs:enumeration value="GENITIVE"/>
                                    <xs:enumeration value="NO_GENITIVE"/>
                                    <xs:enumeration value="UNMARKED"/>
                        </xs:restriction>
            </xs:simpleType>
            <xs:simpleType name="Expletive_Type">
                        <xs:restriction base="xs:string">
                                    <xs:enumeration value="YES"/>
                                    <xs:enumeration value="NO"/>
                                    <xs:enumeration value="det"/>
                                    <xs:enumeration value="der"/>
                        </xs:restriction>
            </xs:simpleType>
            <xs:simpleType name="Definite_Type">
                        <xs:restriction base="xs:string">
                                    <xs:enumeration value="NO"/>
                                    <xs:enumeration value="UNMARKED"/>
                        </xs:restriction>
            </xs:simpleType>
            <xs:simpleType name="NPIndex_Type">
            <xs:annotation>
                                    <xs:documentation> NP and its possible
characteristics
                                    NPIndex and COREF are used express coreference
between NPs                                 in main clauses and in subordinate
constructions
                        </xs:documentation>
            </xs:annotation>
                        <xs:restriction base="xs:string">
                                    <xs:enumeration value="I"/>
                                    <xs:enumeration value="J"/>
                                    <xs:enumeration value="K"/>
                                    <xs:enumeration value="L"/>
                                    <xs:enumeration value="M"/>
                                    <xs:enumeration value="N"/>
                                    <xs:enumeration value="NOINDEX"/>
                        </xs:restriction>
            </xs:simpleType>
            <xs:complexType name="PP_Type">
            <xs:annotation>
              <xs:documentation> PP and its possible characteristics
            </xs:documentation>
            </xs:annotation>
                        <xs:sequence>
                                    <xs:element name="Introducer"
type="Short_String_Type" />
                                    <xs:element name="Clause_Type" type="Clause_T_Type"
/>
                                    <xs:element name="NP_Type" type="NP_T_Type" />
                                    <xs:element name="Control" type="Control_Type"/>
                                    <xs:element name="Coref" type="Coref_Type" />
                                    <xs:element name="Optionalfeature" type="xs:string"
                                                                        minOccurs="0"/>
                        </xs:sequence>
                        <xs:attribute name="Id" type="xs:ID" use="required"/>
            </xs:complexType>
            <xs:simpleType name="Clause_T_Type">
            <xs:annotation>
            <xs:documentation> Types of clause that are prepositional complements
                        </xs:documentation>
            </xs:annotation>
                        <xs:restriction base="xs:string">
                                    <xs:enumeration value="INFINITIVE"/>
                                    <xs:enumeration value="INTERROGATIVE_CLAUSE"/>
                                    <xs:enumeration value="THAT_CLAUSE"/>
                                    <xs:enumeration value="WH_CLAUSE"/>
                                    <xs:enumeration value="NA"/>
                        </xs:restriction>
```

```
            </xs:simpleType>
            <xs:simpleType name="NP_T_Type">
                        <xs:restriction base="xs:string">
                                    <xs:enumeration value="N"/>
                                    <xs:enumeration value="NA"/>
                        </xs:restriction>
            </xs:simpleType>
            <xs:simpleType name="Control_Type">
            <xs:annotation>
              <xs:documentation> Types of control</xs:documentation>
            </xs:annotation>
                        <xs:restriction base="xs:string">
                                    <xs:enumeration value="SUBJECTCONTROL"/>
                                    <xs:enumeration value="OBJECTCONTROL"/>
                                    <xs:enumeration value="INDIRECTOBJECTCONTROL"/>
                                    <xs:enumeration value="PREPOBJECTCONTROL"/>
                                    <xs:enumeration value="RAISING"/>
                                    <xs:enumeration value="NOCONTROL"/>
                                    <xs:enumeration value="WITHOUTCONTROL"/>
                        </xs:restriction>
            </xs:simpleType>
            <xs:simpleType name="Coref_Type">
                        <xs:restriction base="xs:string">
                                    <xs:enumeration value="COI"/>
                                    <xs:enumeration value="COJ"/>
                                    <xs:enumeration value="COK"/>
                                    <xs:enumeration value="CON"/>
                                    <xs:enumeration value="WITHOUTCOREF"/>
                                    <xs:enumeration value="NOCOREF"/>
                        </xs:restriction>
            </xs:simpleType>
            <xs:complexType name="ClauseP_Type">
            <xs:annotation>
               <xs:documentation> Clause and its possible characteristics
                        </xs:documentation>
            </xs:annotation>

                        <xs:sequence>
                                    <xs:element name="Clause_Type" type="Clause_T2_Type"
/>
                                    <xs:element name="Finite" type="YN_Type" />
                                    <xs:element name="Control" type="Control_Type"/>
                                    <xs:element name="Coref" type="Coref_Type" />
                                    <xs:element name="Optionalfeature" type="xs:string"
                                                                          minOccurs="0"/>
                        </xs:sequence>
                        <xs:attribute name="Id" type="xs:ID" use="required"/>
            </xs:complexType>
            <xs:simpleType name="Clause_T2_Type">
            <xs:annotation>
              <xs:documentation> Types of clause that occur as complements (not
                        prepositional complements)</xs:documentation>
               </xs:annotation>
                        <xs:restriction base="xs:string">
                                    <xs:enumeration value="INFINITIVE"/>
                                    <xs:enumeration value="INTERROGATIVE_CLAUSE"/>
                                    <xs:enumeration value="THAT_CLAUSE"/>
                                    <xs:enumeration value="INFINITIVE_NO_INTRO"/>
                        </xs:restriction>
            </xs:simpleType>
</xs:schema>
```