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 6

2 TECHNICAL SPECIFICATIONS 6

2.1 DESCRIPTION OF THE DATA FILES EXTRACTED FROM THE STO DATABASE 6
2.2 DELIVERABLE A: MORPHOLOGY 7
2.3 DELIVERABLE B: SYNTAX 8

3 LEXICON DESCRIPTION 9

3.1 BACKGROUND 9
3.2 CONTENTS OF THE LEXICON 9
3.2.1 Linguistic description: Method and model 9
3.3 COMPOSITION OF THE LEXICON 10
3.4 THE COVERAGE OF THE LEXICON 10
3.5 GENERAL LANGUAGE AND DOMAIN LANGUAGE VOCABULARY 10
3.5.1 Representation of closed word classes 11
3.6 DESCRIPTION OF THE GENERAL LANGUAGE AND DOMAIN LANGUAGE CORPORA 11
3.6.1 General language: corpora and lemma selection 11
3.6.2 Domain languages: corpora and lemma selection 12
3.7 THE ALPHABET OF DANISH 14

4 THE LINGUISTIC CONTENT OF THE LEXICON 15

4.1 ORTHOGRAPHY 15
4.1.1 Spelling and variants in STO 15
4.1.2 Spelling and inflection of new words with foreign origin 15
4.2 THE MORPHOLOGICAL LAYER 16
4.2.1 Treatment of homographs 17
4.2.2 Treatment of spelling variants 17
4.3 MORPHOLOGICAL INFORMATION 18
4.4 INFLECTIONAL BEHAVIOUR 20
4.4.1 Method of description 20
4.5 EXPLANATIONS AND EXAMPLES OF WORD CLASSES 21
4.5.1 Nouns 21
4.5.1.1 Geo-political proper nouns 22
4.5.2 Adjectives 22
4.5.3 Verbs 24

5 FREQUENCY INFORMATION IN STO 25

6 LITERATURE 27

APPENDIX A 30

SPECIFICATIONS FOR MORPHOLOGY EXPORT FROM THE STO LEXICON 30
NOUNS 30

APPENDIX B 33

SPECIFICATIONS FOR MORPHOLOGY EXPORT FROM THE STO LEXICON 33
VERBS 33

APPENDIX C 35

SPECIFICATIONS OF MORPHOLOGY EXPORT FROM THE STO LEXICON 35
ADJECTIVES 35

APPENDIX D 37

SPECIFICATIONS FOR MORPHOLOGY EXPORT FROM THE STO LEXICON 37
OTHER PARTS OF SPEECH 37

APPENDIX E 39

SPECIFICATIONS FOR MORPHOLOGY EXPORT FROM THE STO LEXICON 39
PRONOUNS 39
APPENDIX F 43
SPECIFICATION FOR MORPHOLOGY EXPORT FROM THE STO LEXICON 43
FREQUENCY INFORMATION 43

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

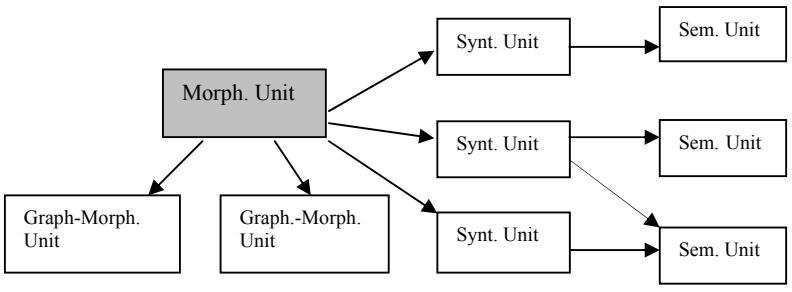
3.2.1 Linguistic description: Method and model

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
blidt	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* whih 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.	