The full-form lexicon

Same word, different versions

DK-CLARIN WP 2.1 Technical Report
Jørg Asmussen, DSL
Draft version of March 5, 2014[1]

Deliverables concerned

D9 Full-form lexicon Development and/or configuration of a full-form lexicon for POS tagging. **Outcome:** Resource with documentation.

D10 Lemmatizer It is considered indispensable that corpus texts need to indicate the lemma form of each inflected word form in the corpus to let the user of the corpus perform more flexible queries. Therefore, it is necessary to either develop or configure a lemmatizer (that may be based on a full-form lexicon or a morphological analyzer). In the context of WP 2.1, a lemmatizer designed as an integral part of a POS tagger is the preferable solution. **Outcome:** Tool with documentation.

D11 POS tagger In order to tag tokens in corpus texts with part-of-speech information, it is necessary to either develop or configure a POS tagger (either based on a full-form lexicon or a morphological analyzer) and a suitable tag set. **Outcome:** Tool with documentation.

[1] A more recent version may be available at:
http://korpus.dsl.dk/clarin/corpus-doc/fullform-lexicon.pdf

Outline of this document

This technical report describes the anatomy of the full-form lexicon that is used for part-of-speech (= POS) tagging. It gives an introduction to material that existed prior to the development of the ePOS tagger (Section 1) and provides an account of how this material was enhanced in order to suit the needs of ePOS tagging. Finally, in Section 2, the ePOS full-form lexicon is described in detail.

1 Enhancing existing material

The input to the full-form lexicon we need for tagging (see Asmussen (2013)) derives from three lexical resources: ONC-Flexion, Flexikon, and – to a certain extent – the PAROLE Corpus itself, cf. Figure .1. ONC-Flexion was derived from existing machine-readable dictionaries in the early 1990s and used as a basis for inflectional information in The Danish Dictionary, DDO. Flexikon was derived from an early version of the DDO around 2000 and used for various purposes in conjunction with the Korpus 2000 website. The PAROLE Corpus, on which the initial language model of the ePOS tagger is based, provides additional lexical entries, however, these entries are not verified and are therefore kept separately in an auxiliary lexicon. At a later point in time new corpus material will be used to enhance the ePOS lexicon that is meant to sever as a primary source for inflectional information in the DDO. The following sections give a more detailed account of the lexical sources of ePOS.

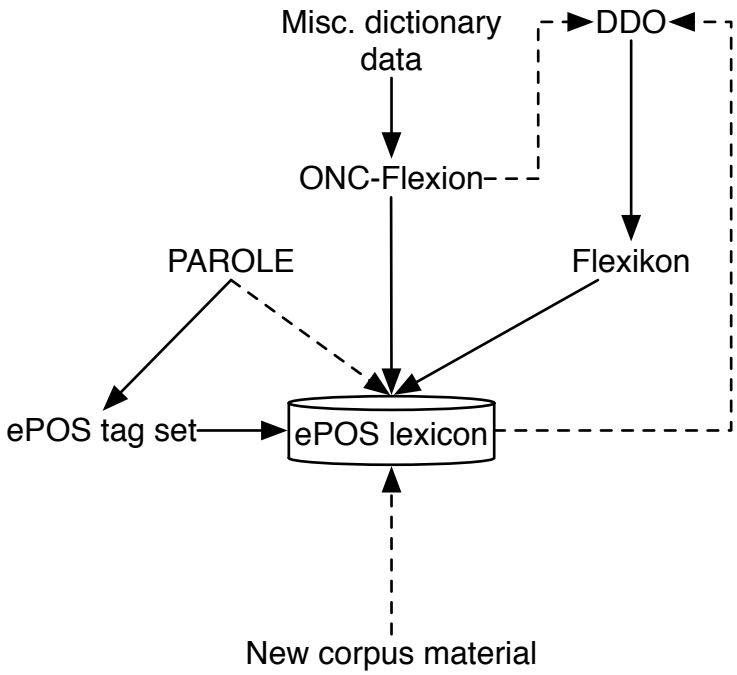

Figure .1: Sources of the ePOS lexicon

1.1 ONC-Flexion

1.1.1 Description

ONC-Flexion[2] is a full-form list with information on parts of speech and inflection for about 80,000 lemmas. ONC-Flexion was originally developed by Ole Norling Christensen in the early 1990s in order to facilitate the process of writing The Danish Dictionary, DDO. ONC-Flexion has since been enhanced by the Korpus 2000 project and a free version of it can be downloaded through the ordnet.dk website. As ONC-Flexion is the most comprehensive and elaborate full-form lexicon of Danish currently freely available, it is used as the major source of the ePOS full-form lexicon.

The lemmas of ONC-Flexion originate from various older sources from the 1980s, and their inflectional forms have been derived from the source information and automatically supplemented in a number of cases. The selection is very wide, and a number of words are hardly relevant (e.g. proper nouns, nonce formations). The majority, however, are words also included in The Danish Dictionary, DDO. In addition, DDO also includes other, particularly newer words that are not in the list.

[2] Free download from:
http://korpus.dsl.dk/e-resurser/boejningsformer_download.php?lang=en

The structure of the full-form list may be illustrated by the following example:

```
*
certifikat
S
2 certifikat
4 certifikats
8 certifikatet
16 certifikatets
32 certifikater
64 certifikaters
128 certifikaterne
256 certifikaternes
```

A new lemma is always preceded by an asterisk (`*`) on a separate line. In the following line the lemma appears in its lemma or base form, in line 3 its part of speech is given. The following part of speech markers occur:

```
S: noun
A: adjective
V: verb
D: adverb
F: abbreviation
K: conjunction
L: onomatopoeic word
O: pronoun
P: proper noun
I: prefix
Æ: preposition
T: numeral
U: interjection
X: unidentified
```

Class `X` comprises words that could not be immediately identified during automatic analysis of the sources. In particular it contains words which usually occur exclusively in fixed expressions with other words, e.g. *badut* (*springe badut*), *bero* (*stille i bero*), *besøgelsestid* (*kende sin besøgelsestid*) or multi word units behaving as a single word, e.g. *au pair*.

In addition to prefixes proper, e.g. *di-*, *eks-*, *fore-*, class `I` also comprises the first elements of compounds, e.g. *forenings-*, *forhandlings-*, *formue-*.

Class `P` comprises proper nouns (i.e. in principle, nouns), but as it is almost impossible to apply usable selection criteria discerning important from unimportant within the class, it is characterized by a degree of coincidence.

Class F reflects primarily an orthographic phenomenon. For all abbreviations, a genitive form with apostrophe *-s* has been generated although many of these seem questionable.

The line containing part of speech is followed by the different orthographic forms which the word may take. The forms are always given in small letters even if capital letters are used in the normally correct orthographic representation. Hyphens, full stops (in abbreviations) and spaces (e.g. *a la*), if any, are also omitted. The omitted information can, however always be derived from the base form.

Each orthographic inflectional form in the list is preceded by a number (separated from the word by a tabulator) indicating which inflectional forms of the lemma the form can be assigned to.

The numbers refer to the bits which have been placed in the following bit pattern:

position	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
value	1	2	4	8	16	32	64	128	256	512	1024	2048	4096	8192	16384	32768

position 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 value 1 2 4 8 16 32 64 128 256 512 1024 2048 4096 8192 16384 32768 If the bits in position 6, 10 and 11 have been entered, the number becomes 64+1024+2048=3136. Different inflectional forms are attached to each position depending on the part of speech. The following tables show the forms attached to the individual positions in the bit pattern for relevant part of speech.

has been built by algorithmically generating full forms based on lemma forms and inflectional information of The Danish Dictionary (DDO). As the structure of inflectional information in the DDO is suboptimal for NLP exploitation a minor group of the automatically generated forms are erroneous, especially composite nouns.

1.1.2 ePOS adaption

ONC-Flexion is based on orthographic forms: each orthographic form is only listed once may have attached several inflectional functions. jaPOS is entirely based on inflectional categories, so the same orthographic forms may be listed under several categories.

ONC-Flexion is based on a slightly different token concept than jaPOS as full stop, hyphen, and apostrophes are considered token characters and not boundaries as is the case in jaPOS. This means that words containing one or more of these characters need special attention. The adaption algorithm identifies these cases in ONC-Flexion and rejects them and puts them on a special list that is checked manually.

2 Anatomy of the ePOS lexicon

3 Inflectional paradigms

3.1 Nouns

3.2 Lexical and inflectional elements

More on how these forms are tagged can be found in Asmussen (2013)

4 Document history

The most recent version of this report can be downloaded from:

- http://korpus.dsl.dk/clarin/corpus-doc/pos-survey.pdf

5 References

Asmussen, J. (2013). Design of the jaPOS tagger. Technical report, DK-CLARIN, korpus.dsl.dk/clarin/corpus-doc/pos-design.pdf.