Design of the ePOS tagger

Making words tell who they are

Technical Report,
Jørg Asmussen, DSL
Final version of May 5, 2015[1]

DK-CLARIN WP 2.1 deliverables concerned

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/pos-design.pdf

Outline of this document

This technical report gives an account of the ePOS tagger that is in part based on Sujit Pal's HMM implementation outlined in Asmussen (2013a).[2] The Danish PAROLE corpus is chosen as source for the language model that the tagger needs in order to work. Even if the quality of the PAROLE corpus is fairly high, it comprises some inconsistencies and mistakes that need to be adjusted before it is viable as a source for a language model. The modifications undertaken in order to enhance the PAROLE corpus are described in Section 1. The concepts and functionality of ePOS as well as the tag set and the construction of the language model are the main topics of Section 2.

1 Modifications of the PAROLE Corpus

The most important considerations on text formats as outlined in Asmussen (2013b) were to keep things as simple as possible. This means that the process of segmenting a text into smaller units should not imply linguistic prior knowledge of any kind. Instead, a mechanical, algorithmic approach is preferred. However, this introduces some intricacies when using PAROLE as a basis for the language model as PAROLE applies a linguistically informed approach to the concepts of sentences and words, and it segments the texts accordingly prior to the POS annotation process. PAROLE thus cannot be used as is as input to a language model that will be applied on material segmented by another, more mechanical approach than PAROLE. Therefore, some modifications of PAROLE were inevitable. These modifications – totaling to nearly 9000 cases – are described in detail in the following sections. The resulting modified version of PAROLE (*PAROLE Version 2*) is freely available upon request.

[2] http://sujitpal.blogspot.com/2008/11/ir-math-in-java-hmm-based-pos.html

1.1 Sentences

PAROLE is subdivided into sentence-like textual units enclosed by <s> and </s> tags. Feeding the language model builder with this kind of textual chunks means that the material to be POS-tagged later on also should resemble that form to a certain extent. This requires a PAROLE type of sentence splitter to be applied prior to (or during) tagging. Punctuation within sentences may to some extent help building a reliable language model during the training phase but must then also be part of the input to be tagged.

The simplest solution would be to work on material without sentence boundary markers but take into account punctuation during training and tagging as this implies a minimum of preprocessing, i.e. just pre-tokenization yielding basic tokens. However, as Sujit Pal's HMM tagger requires the input material during training and tagging to be divided into sentence-like chunks, we end up with a solution where the material is split into individual sentences by some kind of sentence splitting algorithm (baked into the tagger) and where we take into account sentence-internal punctuation.

1.2 Tokens and token boundaries

ePOS is entirely based on the concept of *basic tokens* defined in Asmussen (2013b) and has no linguistic concept of what a word is. So, in ePOS, a word is just a string delimited by characters defined as token boundaries. Token boundaries are either space characters or punctuation characters.[3] These basic tokens need to be tagged in some sensible way even in cases where they do not correspond to linguistic concepts of what a word is. The strictly mechanical token concept has certain implications:

Multiword units: As space characters always are treated as token boundaries there is no concept of multiword units. Each token of such a unit is tagged individually.

Punctuation: Abbreviations, numbers, or hyphenated compounds containing punctuation characters like stops, commas, apostrophes, or hyphens are split into basic tokens at the position of the punctuation character. Each of these basic tokens has to be tagged individually.

Tagging parts of what is normally considered words may in some cases seem weird. However, tagging will most likely show a higher degree of consistency as no linguistic knowledge must be provided nor maintained. The tag set applied to handle these special cases is described in Section 2.

As the PAROLE corpus applies a more linguistically informed token concept, it allows tokens to contain characters that are considered non-token characters in our context.[4] Therefore, the tokens of PAROLE and their tags are converted into

[3] A list of punctuation characters is found in Asmussen (2013c).

[4] Details are listed in the appendix of Keson (1998b).

basic tokens (i.e. the type of tokens defined by DK-CLARIN) prior to using this corpus for building a language model. In detail, all tokens containing space, stop, hyphen, slash, backslash, comma, or apostrophe characters must be converted into their DK-CLARIN equivalents. Table 1 shows the number of words in PAROLE that need to be split.

Character	Count
Period	2488
Space	1556
Hyphen	2831
Comma	150
Apostrophe	412
Slash	141
Colon	8
Brackets	2
Total	7588

Table 1: Number of PAROLE words that are split into basic tokens

The process of splitting such words, which affects approximately 7600 tokens, was carried out manually and semi-automatically.

1.3 Other PAROLE modifications

Apart from re-tagging split words, the following modifications of the PAROLE corpus were carried out:

- Text errors (spelling errors, typos, inflectional errors, etc.) that were tagged `XX` in PAROLE were manually corrected and re-tagged. All 1053 `XX` tags in PAROLE 1.x have thus been converted into meaningful tags in PAROLE 2.0 instead.

- During the process of manually modifying `XX` tags some hundred other tagging errors were identified and corrected.

2 The ePOS tag set for Danish

The tag set applied in the ePOS tagger provides POS and inflectional information, i.e., ideally, for each possible inflectional form of a lemma a corresponding, unambiguous tag is assigned. Tags outside this strictly inflectional scope, e.g. on syntax, semantics, or morphological composition of lemmas, are currently not provided.

As the tagger is trained on the Danish PAROLE Corpus, the tag set of the ePOS tagger will be based on that one used by PAROLE (see Keson (1998a) and Keson (1998b)), however with some modifications, some of them as a consequence of the modified token concept, cf. Section 1.2, some of them for simplification reasons in order to hopefully achieve a better language model. The ePOS tagger utilizes a full-form lexicon that is compiled from the following three resources:

- An existing full-form lexicon derived from an earlier version of The Danish Dictionary: *FLEXIKON*.
- The freely available *FLEXION* lexicon established by Ole Norling-Christensen and others, called *ONCLEX* in the following to better distinguish it from FLEXIKON. [5]
- A supplementary lexicon derived from the tagged PAROLE material.

The ePOS full-form lexicon is described more in-depth in Asmussen (2013d).

2.1 Tag structure

The PAROLE tag set is positional which means that within a tag a certain inflectional marker is always found at a fixed position in a sequence of markers making up the tag. For example, in the case of nouns, the gender marker is always found at position 3, number at position 4, case at position 5, and definiteness at position 8. However, the positional system is dependent on the POS in question: In the case of verbs, which also may carry nominal inflectional markers, definiteness is still at position 8, but gender is at 4, number at 6, and case at 11 (see Keson (1998a) and Keson (1998b)). These varying positions would make it very cumbersome later on to perform corpus queries of the type *find all words that are marked for case = "genitive" no matter what their POS is.* The ePOS tag set therefore favors fixed, POS-independent marker positions.

In contrast to ePOS and PAROLE, the tag set applied by FLEXIKON and ONCLEX is a compact tag set that leaves out non-applicable and implicit information. Hence, it is easy to decode for humans, but may be difficult to formulate complex morphological corpus queries on. Therefore, the tag sets of these sources needs to be converted into a positional one with fixed marker positions independent of the POS in question.[6]

The basic structure of an ePOS tag is:

```
CLASS:nominal:verbal:additional
```

[5] ONCLEX can be downloaded from:
http://korpus.dsl.dk/e-resurser/boejningsformer_download.php?lang=en

[6] The CST tag set used in the CST tagger (see Asmussen (2013a)), which is also based on PAROLE (and seems in part to silently utilize ONCLEX as well), applies compact tags too, cf. the description of the CST tag set at

- http://cst.dk/online/pos_tagger/rapport/bilag/tagset.html.

The same applies to the VISL tag set.

where *CLASS* is a two-character POS classifier comprising a POS indicator (first character) and a sub-classifier. The first colon indicates the boundary between the *CLASS* part and the inflectional part of the tag. Here, *nominal* and *verbal* are strings of markers concerning nominal and verbal morphological information respectively. The *additional* string carries further markers relevant to adjectives, some adverbs, and pronouns. Marker strings have fixed lengths – nominal 4, verbal 2, and additional 4. Each marker in such a string is represented by one *character*. The three groups of morphological markers are separated by colons (**:**). A morphological marker which is irrelevant to a certain paradigm is marked with a dash (**-**) at the respective position(s) of the tag. In cases where inflectional markers are underspecified, this is indicated by a hash sign (**#**) at the respective position(s), meaning *any value of this category*. In certain cases PAROLE applies tags that are underspecified beyond the level of underspecification that the ePOS tag set envisages, that is, they could all be disambiguated by looking at the context in which they occur. Instead of manually disambiguating these tags in the PAROLE corpus prior to using it as a source for the language model of the tagger, they have been adopted by ePOS. In these cases, underspecified markers are marked with a section sign (§). In order to further adapt PAROLE to ePOS, the ePOS-tagged version of PAROLE should be examined and §-underspecifications should be manually disambiguated.[7] The applied markers in ePOS reflect the choices made in PAROLE, cf. its documentation (Keson (1998a) or Keson (1998b)).

Nominal markers

The string of *nominal* markers is of length four, it carries the following markers:

1. **Number** (NUM): *singular* (**s**) or *plural* (**p**)
2. **Definiteness** (DEF): *indefinite* (**i**) or *definite* (**d**)
3. **Case** (CAS): *unmarked* (**u**), *genitive* (**g**), or *fossilized* (**f**), and – for personal pronouns only – *nominative* (**n**) (*accusative* is identical with *unmarked* in these cases and tagged with **u**)
4. **Gender** (GEN): *common* (**c**) or *neuter* (**n**)

Like PAROLE, ePOS considers *gender* an inflectional category – not only of adjectives and verbal participles but for nouns as well, whereas ONCLEX leaves out any explicit information on the gender of a noun and considers this phenomenon as inherent to them.

Verbal markers

The string of *verbal* markers comprises two marker positions:

[7] A future project worthwhile to consider.

1. **Tense** (TMP): *present* (**s**), *past* (**t**)
2. **Voice** (VOC): *active* (**a**), *passive* (**p**)

Additional markers

Finally, *additional* markers constitute a heterogeneous group of the following four markers:

1. **Degree** (DEG, adjectives and some adverbs): *positive* (**p**), *comparative* (**c**), *superlative* (**s**), *absolute superlative* (**a**)
2. **Person** (PER, personal and possessive pronouns): *first* (**1**), *second* (**2**), *third* (**3**)
3. **Reflexiveness** (RFL, personal and possessive pronouns): *yes* (**y**) or *no* (**n**)
4. **Possessor** (POS, possessive pronouns): *singular* (**s**) or *plural* (**p**)

The following structure shows the positions of the tags used in ePOS:

CLASS	*nominal*				*verbal*		*additional*			
	NUM	**DEF**	**CAS**	**GEN**	**TMP**	**VOC**	**DEG**	**PER**	**RFL**	**POS**
	s	i	u	c	s	a	p	1	y	s
	p	d	g	n	t	p	c	2	n	p
			f				s	3		
			n				a			

Table 2: Inflectional markers

All three marker strings are always present in a tag even if some of them are unused (-) or underspecified (# or §). This ensures that it is always straightforward to query on certain marker positions regardless of the POS in question.

The *CLASS* part of the tag comprises two characters. The first character indicates part of speech, the second one may indicate a subclass. If there is no subclass, the second character is a *dash* sign (-). Otherwise, for inflecting words, the subclass always is triggered by a variation of the inflectional paradigm of that particular POS, e.g. participle forms of verbs are characterized by the paradigm `VP:****:*-:----` with inflectional markers (from the table and descriptions above) occurring at the positions marked * whereas finite forms have their inflectional markers according to the paradigm `VF:----:**:----`. Hence verbs come in finite (tagged VF) or in participle (VP) flavor – as well as in a number of other inflectional paradigms, cf. Table 3 in the next section.

2.2 POS markers and subclassifiers in ePOS

2.2.1 Class tags

Table 3 shows the *class tags* used in ePOS, i.e. the symbols used at the first position of the ePOS tags indicating the part-of-speech (Column *POS*) as well as the symbols at the second positions of the ePOS tags giving a potential subclass (Column *Sub.*). The *Paradigm* column shows the structure of the full tag where marker positions with an ⋆ carry inflectional information (denoted by one either character from Table 2 or # or § as discussed in Section 2.1), whereas positions with a dash are unused within the given class. The only exception from this is the `VT` tag marking the past participle form of verbs that has a constant string of inflectional markers (`siu#:t-:----`).

POS		Sub.		Paradigm
V	**Verb**	I	infinitive	`VI:----:-*:----`
		F	finite	`VF:----:**:----`
		M	imperative	`VM:----:--:----`
		G	gerund	`VG:****:--:---`
		P	participle	`VP:****:*-:----`
		T	past part.	`VT:siu#:*-:----`
		D	adv. part.	`VD:----:*-:----`
A	**Adjective**	C	common	`AC:****:--:*---`
		D	adverbial	`AD:----:--:*---`
L	**Numeral**	C	cardinal	`LC:--*-:--:----`
		O	ordinal	`LO:--**:--:----`
N	**Noun**	C	common	`NC:****:--:----`
		P	proper	`NP:****:--:----`
P	**Pronoun**	C	reciprocal	`PC:*-*-:--:----`
		M	demonstrative	`PM:*-**:--:----`
		I	indefinite	`PI:*-**:--:----`
		O	possessive	`PO:*--*:--:-***`
		P	personal	`PP:*-**:--:-**-`
		R	relative	`PR:*-**:--:----`
D	**Adverb**	-		`D-:----:--:*---`
I	**Interjection**	-		`I-:----:--:----`
T	**Preposition**	-		`T-:----:--:----`
C	**Conjunction**	C	coordinating	`CC:----:--:----`
		S	subordinating	`CS:----:--:----`
U	**Unique**	I	inf. marker	`UI:----:--:----`
		S	*som/der*	`US:----:--:----`
E	**Lexical element**	W	word formation	`EW:----:--:----`
M	**Inflectional ending**	N	attached to a noun	`MN:****:--:----`
		V	attached to a verb	`MV:----:**:----`
		A	attached to an adj.	`MA:****:--:*---`
X	**Residual**	S	symbol	`XS:----:--:----`
		F	foreign	`XF:----:--:----`
		Y	tagging error	`XY:----:--:----`

Table 3: POS markers and subclassifiers

As the token concept (cf. Section 1.2) underlying ePOS considers hyphens and apostrophes as token delimiters, we have to deal with special cases of tokens that are not words themselves but parts of words. These may be tagged as *lexical elements*, *inflectional morphemes*, or *word formation elements*. Further details about these types of tokens can be found in Asmussen (2013d).

2.2.2 Lexical elements and inflectional endings

Lexical elements are parts of words that cannot be tagged as regular parts-of-speech. They are often prefixes attached to a word by a hyphen. As such they always play a role in lexical word formation. Lexical elements do not possess any morphological markers, thus the corresponding part of the tag is always set to `----:--:----`. Examples of such elements are

▶	*anti-*	*social*	*adfærd*
	`EW:----:--:----`	`AC:siuc:--:p---`	`NC:siuc:--:----`

▶	*øko-*	*tapas*
	`EW:----:--:----`	`NC:piu#:--:----`

Inflectional endings are grammatical morphemes attached to a noun, verb, or adjective by an apostrophe. Some examples of this are

▶	*cv*	*'er*
	`NC:siun:--:----`	`MN:piu#:--:----`

▶	*vinderen*	*ta*	*'r*	*det*	*hele*
	`NC:sduc:--:----`	`VI:----:-a:----`	`MV:----:sa:----`	`PM:s-un:--:----`	`AC:sdu#:--:p---`

▶	*det*	*go*	*'e*	*vejr*
	`PM:s-un:--:----`	`AC:siuc:--:p---`	`MA:§§u§:--:p---`	`NC:siun:--:----`

Inflectional endings are lexicalized in the full-form lexicon as a special case of lemmas. They can easily be identified as they all start with an @ character.

2.2.3 Word formation elements

Another effect of the token concept is that virtually any POS marker can also occur with the *word formation* subclassifier `W` (In Table 3 this is only listed for the the lexical element `E` that occurs with the `W` subclassifier only). Word formation elements do not possess any morphological markers, thus the corresponding part of the tag is always set to `----:--:----`, e.g. a noun token taking part in word formation is tagged with `NW:----:--:----`. Some examples are listed below.

▶	*sidde-*	*eller*	*sovepladser*
	`VW:----:--:----`	`CC:----:--:----`	`NC:piu#:--:----`

▶	*super-*	*formand*
	`AW:----:--:----`	`NC:siuc:--:----`

►	*planlægnings-,*	*forvaltnings-,*	*og*	*serviceopgaver*
	`NW:----:--:----`	`NW:----:--:----`	`CC:----:--:----`	`NC:piu#:--:----`

►	*den*	*15-*	*årige*	*rocktøs*
	`PM:s-uc:--:----`	`LW:----:--:----`	`AC:sdu#:--:p---`	`NC:siuc:--:----`

►	*fanden-*	*i-*	*voldsk*
	`NM:----:--:----`	`TW:----:--:----`	`AC:siu§:--:p---`

2.2.4 PAROLE's *residual* group in ePOS

Tokens that cannot be identified as a regular part-of-speech are assigned to the *residual* group in PAROLE. This group comprises abbreviations (tagged `XA`), foreign words (`XF`), formulae (`XR`), symbols (`XS`), punctuation (`XP`), and other (`XX`). In ePOS, these subclasses have been reduced to `XS` (symbols) and `XF` (foreign) only, i.e. abbreviations, formulae, symbols, and other have been collapsed into `XS` whereas `XF` is maintained, and `XP` is omitted. The `XY` tag comprises cases where no adequate tag could be assigned, either because the token could not be identified (not in the lexicon) or because the language model could not handle the token in question (even if it is in the lexicon). All tags of the residual group have their morphological part set to `----:--:----`.

3 Document history

A more recent version of this report may be downloaded here:

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

4 References

Asmussen, J. (2013a). Survey of POS taggers. Technical report, DK-CLARIN, korpus.dsl.dk/clarin/corpus-doc/pos-survey.pdf.

Asmussen, J. (2013b). Text formatting. Technical report, DK-CLARIN, korpus.dsl.dk/clarin/corpus-doc/text-format.pdf.

Asmussen, J. (2013c). Text processing. Technical report, DK-CLARIN, korpus.dsl.dk/clarin/corpus-doc/text-processing.pdf.

Asmussen, J. (2013d). The full-form lexicon. Technical report, DK-CLARIN, korpus.dsl.dk/clarin/corpus-doc/pos-design.pdf.

Keson, B. K. (1998a). Documentation of The Danish Morphosyntactically Tagged PAROLE Corpus. Technical report, DSL, korpus.dsl.dk/e-resurser/paroledoc_en.pdf.

Keson, B. K. (1998b). Vejledning til det danske morfosyntaktisk taggede PAROLE-korpus. Technical report, DSL, korpus.dsl.dk/e-resurser/paroledoc_dk.pdf.