TUTORIAL: the CLARIN-DK Workflow Management System

CLARIN’s workflow management system (WMS) combines tools for transformations and annotations of resources into workflows. The WMS is result-centred rather than tool centred. The user specifies a goal (e.g. PoS annotation of a resource) and the WMS computes the ways in which the user’s goal can be reached.

A user of the WMS specifies a goal by choosing a content type from the following list: tokens, sentences, segments, PoS-tags, lemmas, or dependency structures.

Output formats can vary from goal to goal. Some workflows produce data in an XML format (TEI P5), others produce data in a tab separated format (CONLL 2009), a bracketed LISP-like format (Penn Treebank) or as a plain one-item-per-line format, sorted alphabetically or according to frequency.

Steps to follow

To use the WMS, the user must follow these steps:

1. Put one or more text resources in the basket (Danish: ‘kurv’).
2. Press the ‘Tool box’ (‘Værktøjskasse’) button.
3. The system will often seem to report a problem, showing the text
**Access problem**
You have resources in the basket that require that you first accept the Clarin.dk Public License. Click here to read and accept the licence terms.

or, in Danish

**Adgangsproblemer**
Du har ressourcer i kurven som er underlagt Clarin.dk Public License, som du ikke har accepteret. Klik her for at læse og acceptere licensen.

If that is the case, follow the link under ‘Click here to read and accept the licence terms’ (Danish: ‘Klik her for at læse og acceptere licensen’).

Read the licence, mark the box ‘I accept Clarin.dk Public license!’ (Danish: ‘Jeg accepterer Clarin.dk Public licensen!’), and press the ‘Accepter’ button. After 5 seconds the system says ‘Wait a moment…’ During that time the input is analysed.

The system presents a simple form, named ‘Annotation of resources in the basket’ (Danish: ‘Annotation af resurser i kurven’).

(4) If you are logged in, there is only one field to fill in, called ‘Type of content’ (Danish ‘Annotationstyper’). If you are not logged in, you are also asked to fill out your E-mail address. The type of content must be chosen from the drop down list. Not all values can actually be chosen. Starting from a resource in the CLARIN repository it is currently not possible to produce output for the choices ‘text’ (since the input already is text), ‘paragraphs’, and ‘name entities’. All other choices are available.
(5) When you have made your choice, press the ‘Next’ ('Næste’) button. You will see a list of one or more workflows that all are able to produce the type of content that you have chosen, often in several different file formats and combined with several other types of content. If there is more than one workflow, you have to choose one of them before continuing. If in doubt, choose the first workflow in the list. The system tries to organize the workflows in such a way that the ‘best’ appear near the beginning of the list.

(6) Press the ‘Start job’ button.

(7) Keep an eye on your email box. The WMS will send you a notification with a link to the results.

(8) Download the results to your computer.

**Output Examples**
Below is a list of examples of what you can expect as output:

**Tokens**
(TEI P5)

```
<span from="#x00057" xml:id="t37">Item</span>
<span from="#x00059" xml:id="t38">Om</span>
```

**Sentences** (da: ‘Sætninger, før tokenisering’)
(TEI P5)

```
<span from="#x00057" to="#x00076" type="Ssent" xml:id="Z2">Item Om det Tal som faldt aff Himmelen med Lucifero.</span>
```

**Segments** (da: ‘Sætningssegmenter’)
(TEI P5)

```
<span from="#t37" to="#t48" type="Sseg" xml:id="S2" />
```

**PoS tags**
(TEI P5)

```
<span xml:id="P28" from="#t32">TEGN</span>
<span xml:id="P29" from="#t33">ADJ</span>
```
Lemmas
(alphabetic list)
...
1 fælles (1 fælles)
3 følge (1 følge|2 følgende)
1 før (1 før)
1 først (1 første)
3 gensidig (3 gensidig)
3 give (1 give|1 gives|1 givet)
2 gælde (2 gælder)
...

(frequency list)
36 . (36 .)
35 , (35 ,)
33 af (33 af)
27 for (27 for)
22 og (22 og)
17 i (17 i)
15 til (15 til)
14 en (7 en|7 et)
14 inden (14 inden)
12 arkitekturområde (12 arkitekturområdet)
10 at (10 at)
10 den (3 de|7 den)
9 - (9 -)
9 1 (9 1)
9 uddannelse (4 uddannelse|5 uddannelsen)
9 udvalg (5 udvalg|1 udvalget|3 udvalgets)
8 / (8 /)
8 Kommission (8 Kommissionen)
8 det (8 det)
8 medlem (3 medlem|4 medlemmer|1 medlems)
8 rådgive (1 rådgive|6 rådgivende|1 rådgiver)
...

Syntax (dependency structures)

(CONLL)

1 Foto _ Foto NOUN NOUN _ _ -1 2 _ nobj _ _
2 udlånt _ udlåne VERB VERB _ _ -1 0 _ ROOT _ _
3 af _ af ADP ADP _ _ -1 2 _ pobj _ _
4 søsteren _ søster NOUN NOUN _ _ -1 3 _ nobj _ _
5 Petra _ petra NOUN NOUN _ _ -1 4 _ possd _ _
6 , _ _ X X _ _ -1 5 _ pnct _ _
7 gift _ gift ADJ ADJ _ _ -1 8 _ mod _ _
8 Juul _ juulen NOUN NOUN _ _ -1 5 _ title _ _
9 _ _ _ _ _ _ _ _ _ _ _ _ _
List of tools

**TEI P5 Segmenter**

Produces segment annotations. The output from this tool is important for some other tools.

**TEI P5 Tokenizer**

Produces tokens. The output from this tool is important for some other tools.

**POS-Tagging openNLP**

Open source POS tagger, supports multiple languages. We have selected Danish and English for the CLARIN–DK platform. This POS-tagger assigns word classes (NOUN, VERB, ADJ, etc.) to tokens in the input text, but no information about the particular inflections of the tokens.

**CST Lemmatiser**

The lemmatizer derives the dictionary look-up form (lemma) of words.

A lemmatizer is a valuable human language technology tool to improve precision and recall in an information retrieval setting. For example, lemmatization makes it possible to match a query in one morphological form with a word in a document in another morphological form. Lemmatizers can also be used in lexicography to find new words in text material, including the words' frequency of use. Other applications are creation of index lists for book indexes as well as key word lists. Lemmatization is the process of reducing a word to its base form, normally the dictionary look-up form (lemma) of the word.
**CONLL Converter**
This tool merges several annotation layers represented in the TEI P5 format into a single CoNLL 2009 formatted file.

**Brill’s PosTagger**
CST’s POS-tagger is an expanded version of Brill-tagger, with add-ons for handling of XML and for better handling of words with capital letters e.g. in headlines. In CLARIN-DK Brill’s PoS-tagger supports the languages Danish and English. **This POS-tagger assigns word classes and information about inflection to tokens in the input text (N_INDEF_SING, V_PARTC_PAST, ADJ).**

**Bohnet’s parser**
Bohnet’s parser is a syntactic dependency parser.