Text formatting

Bringing corpus texts into good shape
and enabling flexible annotation of them

DK-CLARIN WP 2.1 Technical Report.
Jørg Asmussen, DSL, with input from other WP 2 members
Final version of June 15, 2011[1]

Deliverables concerned

D 2: Tokenizer This document provides the definition of the token concept used throughout WP 2.1. The token concept has profound implications on the design of the tokenizer tool (Asmussen (2011a)) and the POS-tagger (Asmussen (2011b)) applied in WP 2.1.

D 13: TEI transducer The original plan for WP 2.1 was based on the assumption that the repository of potential corpus texts – the corpus text bank – would be a relational database. In order to make interchange of texts easy and in order to make them fit into the intended resource repository of DK-CLARIN, the development of a transducer that could reshape the texts and metadata stored in the relational corpus text database to valid TEI XML seemed necessary. However, during the course of the project, it became clear that the text bank itself should be implemented as an XML database so that the texts could be stored in their final TEI XML format. Therefore, the task of developing a transducer became a task of defining an appropriate subset of TEI in order to suit the metadata and text format needs of DK-CLARIN.

[1] The most recent version can be downloaded from:
http://korpus.dsl.dk/clarin/corpus-doc/text-format.pdf.

Outline of this document

This technical report gives an overview of the *general* principles of text formatting and word level markup for corpus texts. For more *specific* information, consult the documentation of applications and procedures for segmenting and annotating text, see Asmussen (2011a) and Asmussen (2011b).

1 Basic considerations

1.1 Motivation

The main motivation of defining a *general text format* is to establish a joint basis for all tools that operate on CTB texts. Thus, tools do not need to be configured for a multitude of formats which means that they will be easier and less error-prone to develop and maintain.

1.2 Format requirements

1. The format has to be expressed by means of TEI P5.
2. Annotations should not interfere with the basic format of the text proper.
3. The basic format of the text proper should not be biased by interpretations.
4. It must be possible to annotate one single text with various (possibly mutually exclusive) types of annotations, each type appearing as a group of annotations that conceptually belong together.

5. Each annotation in an annotation group must be able to refer to either the text proper or to another annotation group which means that layers of annotations, i.e. annotations on annotations, become feasible.

6. It should be possible to store annotations separate from the text proper.

7. Several versions of the text proper should be avoided.

1.3 Consequences

Pre-tokens: The text has to be mechanically segmented into rather primitive tokens which we call *pre-tokens* in the following as they do not reflect any linguistic word conceptualizations.

Reference: It must be possible to unequivocally refer to these tokens.

Transformation: A generalized, multi-purpose format that needs to be transformed in order to be legible for humans which means that specific viewers and editors must be developed in order to interact with the text.

2 Formatting text

2.1 A source sample to be formatted

The following snippet shows a paragraph taken from the DDOC[2] source files:

```
<p><s>To kendte russiske historikere Andronik Mirganjan og Igor
Klamkin tror ikke, at Rusland kan udvikles uden en
"jernnæve".</s></p>
```
[3]

A text version like this one is called the *source version* of a text. The source version of a text must comply with the TEI P5 specifications in order to be formatted. If it does not, it must be converted into TEI P5 prior to further formatting. The excerpt above conforms to TEI.

2.2 Bad: Formatting against the requirements

In the DK-PAROLE Corpus, cf. Keson (1998), this same paragraph/sentence is formatted like this:

[2] DDOC = Corpus of The Danish Dictionary, cf. Norling-Christensen and Asmussen (1998).

[3] Original source: Leon Nikulin: Jeltsins skæbnetime, Det Fri Aktuelt, 1.12.1992, p. 14. Actually, the original paragraph is longer than the single sentence reproduced here.

```
<p>
  <s>
    <W lemma="to" msd="AC---U=--">To</W>
    <W lemma="kendt" msd="ANP[CN]PU=[DI]U">kendte</W>
    <W lemma="russisk" msd="ANP[CN]PU=[DI]U">russiske</W>
    <W lemma="historiker" msd="NCCPU==I">historikere</W>
    <W lemma="Andronik" msd="NP--U==-">Andronik</W>
    <W lemma="Mirganjan" msd="NP--U==-">Mirganjan</W>
    <W lemma="og" msd="CC">og</W>
    <W lemma="Igor" msd="NP--U==-">Igor</W>
    <W lemma="Klamkin" msd="NP--U==-">Klamkin</W>
    <W lemma="tro" msd="VADR=----A-">tror</W>
    <W lemma="ikke" msd="RGU">ikke</W>
    <W lemma="," msd="XP">,</W>
    <W lemma="at" msd="CS">at</W>
    <W lemma="Rusland" msd="NP--U==-">Rusland</W>
    <W lemma="kunne" msd="VADR=----A-">kan</W>
    <W lemma="udvikle" msd="VAF-=----P-">udvikles</W>
    <W lemma="uden" msd="SP">uden</W>
    <W lemma="en" msd="PI-CSU--U">en</W>
    <W lemma="&quot;" msd="XP">"</W>
    <W lemma="jernnæve" msd="NCCSU==I">jernnæve</W>
    <W lemma="&quot;" msd="XP">"</W>
    <W lemma="." msd="XP">.</W>
  </s>
</p>
```

Even if the format is easy to decode, at least for humans, it has certain shortcomings running counter to the requirements defined in Section 1.2 above:

1. The format is not expressed by means of TEI P5 as <W> is not a legal element in TEI P5 (whereas <w> would be, but *msd* is not a legal attribute of <w>).

2. Annotations interfere with the format of the text proper (as attributes of the <W>-element).

3. The text format is affected by interpretation: Punctuation characters are considered as words that again carry a lemma tag and a morphosyntactic tag.

4. New annotation layers can hardly be added without further interfering with the already existing format (e.g. by adding further attributes to the <W> element).

5. It is not possible to refer to the basic tokens of the text.

6. Annotations cannot easily be separated from the text proper.

7. Other interpretations of the text expressed by alternative annotations may require new versions of the text.

2.3 Good: Formatting according to the requirements

2.3.1 From source version to base format

Two of the consequences emerging from the requirements are that the text has to be mechanically segmented into *basic tokens* and that it must be possible to unequivocally refer to these tokens, cf. Section 1.3. Mechanical text segmentation, or pre-tokenization, is carried out by certain textual surface items, i.e. characters, only. For segmentation purposes characters fall into three categories:

- Letters and numbers, i.e. alpha-numeric characters
- Whitespace characters
- Punctuation characters

Continuous sequences of alpha-numeric characters are considered 'words' even if these segments are not necessarily in accordance with a linguistic definition of a word. Linguistic interpretations are deliberately avoided at this point. 'Words' are put into `<w>` elements.

Whitespace and punctuation is put into `<c>` elements – character by character – that can be of *type* "s" (space) or "p" (punctuation). The non-obligatory *subtype* attribute may specify some other characteristics of the character in question, e.g. the length of a whitespace. Specifications of the possible inventory of the subtype attribute are not given before it turns out that this attribute is really needed. Standard space characters (ASCII 32) are not explicitly denoted in the `<c>` elements (i.e. they remain empty) whereas other whitespace characters such as tabs (coded as `&#x9;`) can be given in the element.

`<w>` and `<c>` elements are the smallest segments (i.e. *basic tokens*) of a text. Each of them carries a unique *xml:id* that allows referencing to it from elsewhere.[4] The source example given in Section 2.1 would look like this after segmentation:

```
<p>
  <s>
    <w xml:id="x002">To</w>
    <c xml:id="x003" type="s"/>
    <w xml:id="x004">kendte</w>
    <c xml:id="x005" type="s"/>
    <w xml:id="x006">russiske</w>
    <c xml:id="x007" type="s"/>
    <w xml:id="x008">historikere</w>
    <c xml:id="x009" type="s"/>
    <w xml:id="x010">Andronik</w>
    <c xml:id="x011" type="s"/>
```

[4] Assigning IDs requires some sort of control that every ID is unique.

```
      <w xml:id="x012">Mirganjan</w>
      <c xml:id="x013" type="s"/>
      <w xml:id="x014">og</w>
      <c xml:id="x015" type="s"/>
      <w xml:id="x016">Igor</w>
      <c xml:id="x017" type="s"/>
      <w xml:id="x018">Klamkin</w>
      <c xml:id="x019" type="s"/>
      <w xml:id="x020">tror</w>
      <c xml:id="x021" type="s"/>
      <w xml:id="x022">ikke</w>
      <c xml:id="x023" type="p">,</c>
      <c xml:id="x024" type="s"/>
      <w xml:id="x025">at</w>
      <c xml:id="x026" type="s"/>
      <w xml:id="x027">Rusland</w>
      <c xml:id="x028" type="s"/>
      <w xml:id="x029">kan</w>
      <c xml:id="x030" type="s"/>
      <w xml:id="x031">udvikles</w>
      <c xml:id="x032" type="s"/>
      <w xml:id="x033">uden</w>
      <c xml:id="x034" type="s"/>
      <w xml:id="x035">en</w>
      <c xml:id="x036" type="s"/>
      <c xml:id="x037" type="p">"</c>
      <w xml:id="x038">jernnæve</w>
      <c xml:id="x039" type="p">"</c>
      <c xml:id="x040" type="p">.</c>
    </s>
  </p>
```

This formatted version of the source text is called the text's *base format*. The base format is the standard input format for all tools like tokenizers, sentence splitters, lemmatizers, and taggers of all kinds, see the motivation for a fixed text format in Section 1.1.

As can be seen, markup above <w> and <c> level that is already present in the source version text, may be kept as long as the source version complies with the TEI specifications. In this case, the <p> and <s> tags were kept; <p> tags may carry an *xml:lang* attribute that indicates the language of the paragraph by using a value from the languageId value set described in Asmussen et al. (2011). Even though tags other than <c>, <w>, <s>, and <p> may be used as long as they are TEI-compliant,[5] this type of markup should be avoided and added as span groups instead, see the following section.

[5] Among such elements is <lb> (line break) whereas the <hl> element, which was introduced by some other WP 2 projects, is not allowed in the body of a text. See also Section 2.3.4 on further examples.

2.3.2 Annotations

Annotations are given separately from the base format version of the text by a number of elements enclosed in <spanGrp> elements. The elements contain the annotations themselves that are either attached to one single basic token or a number of continuous basic tokens. Attachment is achieved by referencing the *xml:id* units from the obligatory *from* attribute of the element and – in case continuous basic tokens are referenced and not only a single one – the facultative *to* attribute. Every <spanGrp> contains one type of annotations only. The *ana* attribute of the <spanGrp> element refers to the application or method that has produced the annotations, listed in the <appInfo> element of the header. Some annotation examples follow.

Sentences In the base format version given in Section 2.3.1 <p> and <s> tags from the source version were kept as independent tags as they occurred above the level of the basic tokens and met the TEI specifications. The <p> tags are an obligatory part of the structure: Raw text as well as <w> and <c> elements must be encapsulated by <p> elements or equivalent elements, e.g. <ab>. However, the <s> tags could alternatively be expressed as <spanGrp> annotations. The following example shows how sentences can be tagged in this alternative way making <s> tags in the the base format version unnecessary.

```
<spanGrp ana="#sentences">
  <span from="#x002" to="#x040">s</span>
</spanGrp>
```

Lemmas The following example shows what the PAROLE lemma annotation expressed by the *lemma* attributes as shown in Section 2.2 looks like when expressed by the <spanGrp> annotation.

```
<spanGrp ana="#paroleLemma">
  <span from="#x002">to</span>
  <span from="#x004">kendt</span>
  <span from="#x006">russisk</span>
  <span from="#x008">historiker</span>
  <span from="#x010">Andronik</span>
  <span from="#x012">Mirganjan</span>
  <span from="#x014">og</span>
  <span from="#x016">Igor</span>
  <span from="#x018">Klamkin</span>
  <span from="#x020">tro</span>
  <span from="#x022">ikke</span>
  <span from="#x023">,</span>
  <span from="#x025">at</span>
  <span from="#x027">Rusland</span>
  <span from="#x029">kunne</span>
```

```
  <span from="#x031">udvikle</span>
  <span from="#x033">uden</span>
  <span from="#x035">en</span>
  <span from="#x037">"</span>
  <span from="#x038">jernnæve</span>
  <span from="#x039">"</span>
  <span from="#x040">.</span>
</spanGrp>
```

The linguistic interpretation expressed by the PAROLE lemma annotation is exactly the same as in the example shown in Section 2.2, including that punctuation characters are treated as lemmas, but this interpretation no longer imposes a certain formatting on the base format of the text. Base format and interpretation are kept apart.

POS and inflection In the same manner, the morphosyntactic annotation of the PAROLE corpus can be expressed by a `<spanGrp>`:

```
<spanGrp ana="#paroleMsd">
  <span from="#x002">AC---U=--</span>
  <span from="#x004">ANP[CN]PU=[DI]U</span>
  <span from="#x006">ANP[CN]PU=[DI]U</span>
  <span from="#x008">NCCPU==I</span>
  <span from="#x010">NP--U==-</span>
  <span from="#x012">NP--U==-</span>
  <span from="#x014">CC</span>
  <span from="#x016">NP--U==-</span>
  <span from="#x018">NP--U==-</span>
  <span from="#x020">VADR=----A-</span>
  <span from="#x022">RGU</span>
  <span from="#x023">XP</span>
  <span from="#x025">CS</span>
  <span from="#x027">NP--U==-</span>
  <span from="#x029">VADR=----A-</span>
  <span from="#x031">VAF-=----P-</span>
  <span from="#x033">SP</span>
  <span from="#x035">PI-CSU--U</span>
  <span from="#x037">XP</span>
  <span from="#x038">NCCSU==I</span>
  <span from="#x039">XP</span>
  <span from="#x040">XP</span>
</spanGrp>
```

Again, punctuation characters are treated as independent units carrying their own morphosyntactic annotation (“XP”).

Alternative POS markup If the morphosyntactic PAROLE annotation is considered inadequate for certain purposes, a new annotation group with another tag set and another treatment of punctuation easily can be added, for example:

```
<spanGrp ana="#parolePOS">
  <span from="#x002">NUM</span>
  <span from="#x004">ADJ</span>
  <span from="#x006">ADJ</span>
  <span from="#x008">S</span>
  <span from="#x010">S/span>
  <span from="#x012">S</span>
  <span from="#x014">KON</span>
  <span from="#x016">S</span>
  <span from="#x018">S</span>
  <span from="#x020">V</span>
  <span from="#x022">ADV</span>
  <span from="#x025">SUB</span>
  <span from="#x027">S</span>
  <span from="#x029">V</span>
  <span from="#x031">V</span>
  <span from="#x033">PRP</span>
  <span from="#x035">ART</span>
  <span from="#x038">S</span>
</spanGrp>
```

Names, manually annotated In the same manner e.g. names could be marked up, for example as result of a manual procedure:

```
<spanGrp ana="#paroleNames">
  <span from="#x010" to="#x012">person</span>
  <span from="#x016" to="#x018">person</span>
  <span from="#x027">place</span>
</spanGrp>
```

2.3.3 Putting base format and annotation layers together

The base format version from Section 2.3.1 and all annotation groups are structurally combined as shown in the following sketch. The text in base format is enclosed by <body> tags whereas the <spanGrp> elements are siblings of the <body> element, following it in arbitrary order:

```
<text>
  <body>[6]
    - Text in base format (with obligatory paragraph markup)
  </body>
    - <spanGrp> with sentence markup[7]
    - <spanGrp> with lemma annotations
    - <spanGrp> with POS and inflectional annotations
    - <spanGrp> with alternative POS markup
    - <spanGrp> with name annotations
```

[6] TEI expects the text to be subdivided into front matter, text body, and back matter. For corpus texts, a subdivision of this kind is unnecessary. However, TEI demands at least the <body> subdivision. Therefore, all CTB <text> elements contain one single <body> element encapsulating the text body.

[7] If the <s> markup is not already contained in the base format version of the text.

```
</text>
```

2.3.4 Additional information in the base version

According to TEI 5, only a few elements may occur as siblings to the <w> and <c> elements. The use of such elements to give additional textual or graphical formatting information should be generally avoided. This type of information should be placed in <spanGrp> elements if it cannot be entirely eliminated.[8]

However, as putting additional information into <spanGrp> elements may complicate the process of converting text from original versions to the base version, some exceptions are the following tags which occur in some forum texts gathered in WP 2.1:

<quote> may occur as sibling of <w> and <c> tags and may embed them as well, i.e. have them as children. This tag is used for surrounding text material that is quoted in forum posts; in these cases it always carries the *type* attribute 'forum'.

<add> may occur as sibling of <w> and <c> tags but cannot contain them. This element gives additional information on extra-textual resources like images (mandatory *type* attribute is 'img', mandatory *source* attribute is the URI of the resource) or pointers (*type* is 'url', *source* as before), or video (*type* 'video', *source* as before).

2.3.5 What happens to the source version of a text?

When converting a TEI P5 source version of a text into base format, all information is kept either as additional markup in the base format version like the <p> and <s> markup in the example shown in Section 2.3.1 or as independent span groups as shown in the sentences example in Section 2.3.2. As all necessary textual and extra-textual information contained in the source version can be expressed in base format in conjunction with a number of span groups, the source version proper becomes obsolete, and thus is not kept as a member of the CTB files. However, it should be stored independently in some location that can be referenced from the CTB header through one of the <idno> elements within <biblStruct>, see Asmussen et al. (2011). The same applies to other source versions like URLs from which the TEI P5 base versions may have been derived. In the case of the WP 2.1 corpus project, all original texts are stored in the text file repository in the /Data volume on the server ja-korpus.dsl.lan – the same server where the eXist-db is installed, see Asmussen (2011c).

[8] As concerns linguistic corpus texts, layout information is dispensable in most cases and therefore can be removed.

2.3.6 Format requirements revisited

So far, the format requirements 1 – 4 and 7 have been highlighted by the example given above in Section 1.2. Regarding requirement 5, it has been shown how an annotation group refers to the text proper, namely through *from* and optional *to* attributes referencing the *xml:id* attributes of the basic textual units. What has not been shown yet is how to layer annotation groups by letting them reference other annotation groups. This will be part of the following examples section. Requirement 6, the possibility of storing annotations separate from the text proper, may be illustrated in a separate document.

2.4 Example

2.4.1 Tokenization and layers of annotations

The base format of a text may to some extent resemble a tokenized version of it. However, 'real' tokenization normally requires a certain amount of language-specific linguistic knowledge on how to identify words, but the segmentation procedure applied to the source version of a text in order to transform it into base format does not possess this kind of knowledge; in fact, the segmentation procedure is entirely ignorant on delicate linguistic considerations. This means, that in some cases it may be desirable to apply a more intelligent tokenization procedure in addition to the mere segmentation of the source text as the following example shows.

Source version To keep this example as simple as possible, markup above the level of the basic textual units is kept to a minimum, i.e. the obligatory <p> tags:

```
<p>De staar over for et PROBLEM i dag.</p>
```

Base format The source version is converted into base format:

```
<p>
  <w xml:id="y01">De</w>
  <c xml:id="y02" type="s"/>
  <w xml:id="y03">staar</w>
  <c xml:id="y04" type="s"/>
  <w xml:id="y05">over</w>
  <c xml:id="y06" type="s"/>
  <w xml:id="y07">for</w>
  <c xml:id="y08" type="s"/>
  <w xml:id="y09">et</w>
  <c xml:id="y10" type="s"/>
  <w xml:id="y11">PROBLEM</w>
  <c xml:id="y12" type="s"/>
```

```
  <w xml:id="y13">i</w>
  <c xml:id="y14" type="s"/>
  <w xml:id="y15">dag</w>
  <c xml:id="y16" type="p">.</c>
</p>
```

Tokenization and regularization The linguistically ignorant segmentation mechanism that converts the source version into base format, treats the two word pairs *over for* and *i dag* as four separate words even if each pair reasonably may be considered as one single word, once in a while also with substandard spelling *overfor* and *idag*. In order to express this linguistically enlightened view on which textual units are to be treated as tokens, a token annotation layer is introduced as a span group. The tokenization algorithm applied furthermore regularizes the spelling of words according to some predefined norm of some kind, in the present case *De* is regularized as *de*, *staar* as *står*, and *PROBLEM* as *problem*. The `<span>` elements of this span group all carry an additional *xml:id* attribute giving each `<span>` a unique ID which can be referenced from elsewhere, i.e. from other annotation groups.

```
<spanGrp ana="#tokenRegular">
  <span xml:id="t1" from="#y01">de</span>
  <span xml:id="t2" from="#y03">står</span>
  <span xml:id="t3" from="#y05" to="#y07">over for</span>
  <span xml:id="t4" from="#y09">et</span>
  <span xml:id="t5" from="#y11">problem</span>
  <span xml:id="t6" from="#y13" to="#y15">i dag</span>
</spanGrp>
```

Lemmatization The following lemma annotations no longer address the basic textual units but instead the `<span>` elements of the annotation group above:

```
<spanGrp ana="#lemma">
  <span from="#t1">de</span>
  <span from="#t2">stå</span>
  <span from="#t3">over for</span>
  <span from="#t4">en</span>
  <span from="#t5">problem</span>
  <span from="#t6">i dag</span>
</spanGrp>
```

POS annotation Finally, the following POS annotations address the same token annotation layer as does the lemma annotation group:

```
<spanGrp ana="#lemma">
  <span from="#t1">PRON</span>
  <span from="#t2">V</span>
  <span from="#t3">PRP</span>
  <span from="#t4">ART</span>
  <span from="#t5">S</span>
  <span from="#t6">ADV</span>
</spanGrp>
```

3 Further reading

- The tag set applied in WP 2.1 is described in

4 Document history

The most recent version may contain some minor fixes that are not explicitly listed here.

Current version (June 15, 2011)

- Minor error fixes.

Earlier versions (most recent changes first)

- Pre-final version, adjusted to the outcome of discussions in the WP2 group on 20 May 2009. Each sub-WP is encouraged to experiment with the format specifications and report on experiences. WP 2.1 starts experimental conversion of text material to the format specified in this documentation. WP 2.1 and WP 2.3 will discuss WP 2.3's conversion. Minor modifications of the text format specifications may occur, therefore each WP2.x should contact Jørg Asmussen before any final conversion tasks are carried out.
- A more clear definition of a text's "base version".
- First version for discussion.

5 References

Asmussen, J. (2011a). CTB web-services. Technical report, DK-CLARIN, korpus.dsl.dk/clarin/corpus-doc/ctb-webservices.pdf.

Asmussen, J. (2011b). Design of the jaPOS tagger. Technical report, DK-CLARIN, korpus.dsl.dk/clarin/corpus-doc/pos-design.pdf.

Asmussen, J. (2011c). Textbank software. Technical report, DK-CLARIN, korpus.dsl.dk/clarin/corpus-doc/textbank-software.pdf.

Asmussen, J. et al. (2011). Text metadata. Technical report, DK-CLARIN, korpus.dsl.dk/clarin/corpus-doc/text-header.pdf.

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

Norling-Christensen, O. and Asmussen, J. (1998). The Corpus of The Danish Dictionary. *Lexikos. Afrilex Series*, 8:223–242.