

# Workpackage 2 - Specification of a target format for the linguistic annotation, Version 3

Lothar Lemnitzer  
Universität Tübingen

October 2006

## 1 Introduction

The aim of this paper is to define a common target format for the linguistic annotation of the learning objects which have been collected in WP1. The target format will be used as the input for the linguistic tools which we are going to build. These tools should, as far as it is possible, be language independent. They should at least accept input from all languages. It is therefore required that the annotated learning objects of all languages conform to the same document type definition. It is the aim of these guidelines to ensure this uniformity.

The target format should meet the following requirements:

- it should allow us to add the linguistic annotation on top of the Base XML format. The target format must therefore be an extension of the Base XML format;
- it should be as close as possible to the xcesAna DTD which defines the linguistic annotation part of the Corpus Encoding Standard;
- it should allow for the annotation of the base form, the morphological features, and the part of speech of tokens (which are the obligatory features) and for the annotation of chunks and named entities (which are optional features);
- we agreed that the annotation of our learning objects will be inline. The xcesAna DTD, from which we derived this DTD, recommends the use of stand-off annotation and provides a set of attributes which support this kind of annotation (cf. xcesAna documentation, Overview). It is however still possible to use the DTD for inline annotation. We therefore decided to get rid of all the attributes which serve the only purpose of enabling stand-off annotation;
- The target format has to be extended to allow for the annotation of keywords and definitions;
- The original text should be easily recoverable from the annotated document.

## 2 Status of this DTD

The DTD is derived from the xcesAna DTD, revision 1.11 of 8 May 1996. For the design of our annotation model, we also refer to the Polish adaptation and the German adaptation of this DTD. Since some changes are substantial, however, we will call this DTD *LT4ELAna*.

There will be two versions of this DTD:

- One version with linguistic and LT4EL specific annotation. We call this DTD *LT4ELAnaProject*;
- One version with linguistic, but without the LT4EL specific annotation. We call this DTD *LT4ELAnaRelease*, because this is the DTD which we will need once the project results will be released.

Note that the original xcesAna DTD has been written for use with large and multi-purpose data collections. We have to annotate relatively small and single-purpose corpora. Some of the features are therefore not necessary to provide and are therefore left out.

### 3 Major changes to the xcesAna annotation model

1. We decided to use inline annotation instead of standoff annotation (see above). We will therefore remove the features which are only used for cross-referencing various related documents. This affects mainly the *from* and *to* attributes of the elements which surround sequences of text.
2. For the same reason (i.e. supporting standoff annotation), the term *chunk* is used in the xcesAna DTD with a non-linguistic meaning. In the XCES document model, the term *chunk* signifies any sequence of text which might be subject to further processing or annotation. In contrast to this, we use the term *chunk* to signify, broadly speaking, a phrase which does not include phrases of the same phrase type. For our annotation model, which should allow for partially parsed output, this means that we will remove the <chunk> element from the text structural hierarchy and move it to the hierarchy of linguistically motivated elements (see below).
3. The elements which are legacy of the XMLBase model have to be integrated to the DTD.

## 4 The hierarchical layers

### 4.1 The linguistic elements

If we remove the *chunk* layer as part of the text structure, we will get the following hierarchy of linguistic objects:

- The document consists of a header (optionally) and a sequence of paragraphs (which are either titles or text paragraphs). Minimum is one paragraph.
- The *paragraph* consists of a of a sequence of sentences.
- The *sentence* consists of a sequence of chunks and tokens.
- The *chunk* consists of a sequence of chunks and/or tokens.
- The *token* is the elementary annotation unit.
- 

In short, the markup of paragraphs, sentences, and tokens is obligatory, the markup of chunks is optional.

### 4.2 The XMLBase elements

The layout information which is inherited from the XMLBase documents should be integrated into the target document mode and thus be available for further processing. We use a *rend* attribute on the level of the token to mark this layout information.

### 4.3 LT4EL specific elements

For further processing of the documents with our NLP tools, for training, testing and validation, we have to encode two LT4EL specific types of data: keyword and definition. The latter is divided into *defined term* (definiendum) and *defining text* (definiens). Since the defined word and the defining text can be discontinuous (see example 2), the two parts of the definition will be two different elements which will be linked through the ID/IDREF mechanism

**example 2:** Now let us define *part of speech*. This term signifies one of the traditional grammatical categories, like *noun*, *verb*, *adjective*, etc.

We will not use this kind of annotation for the running system, though. We will therefore build a *project* version of the DTD and a *release* version of it to be used with the running system. The project version allows for the annotation of these data structures.

## 4.4 Global attributes

We will only slightly change the list of attributes. In particular, we will remove the *n* attribute which is only used to handle standoff annotation.

Second, we strongly recommend each partner to use an identifier with each linguistic element. The use of IDs helps us to identify the exact place(s) where a keyword or defined term occurs (first). We therefore propose that all partner use a scheme of identifiers which start with a letter indicating the type of linguistic element, followed by a sequential number (e.g., *p11* = the eleventh paragraph; *s8* = the eight sentence; *c2* = the second chunk; *t89* = the eighty-ninth token).

We will keep the global attributes *type* (which is a misleading name, since it is supposed to be the place to specify the language of an element's textual content) and *wsd* (which specifies the writing system for the element's textual content). Both attributes are optional (= #IMPLIED), so that you can and should use them only in case where the language / writing system deviates from the convention (e.g. in the case of an English quote in an otherwise Romanian text).

## 4.5 The root element

The name of the root element will be LT4ELAna.

The root element will have a list of *par* (paragraph) elements as content. The root element has as one single attribute, *version*, which reflects to which version of this DTD the document conforms. This attribute should also reflect whether the document conforms to the *project* or to the *release* version of the DTD.

## 4.6 The linguistic elements

The names, attributes, and content models of the linguistic elements are as follows:

- *par*. No text should appear outside a paragraph element. This implies that titles, captions etc. should also be treated as paragraphs. A paragraph contains a sequence (one element minimally) of tokens and sentences.
- *s*, the sentence. According to the xcesAna dtd, the sentence contains #PCDATA, tokens, and other, embedded sentences. We would like to make the following changes here:
  - introduce chunks as parts of sentences. The use of chunks, however, is optional.
  - make the content model stricter. We will not allow the use of textual data (#PCDATA) outside tokens.
  - remove the *next* and *previous* attributes from the attributes list unless anyone needs them.
  - no sentences within sentences are allowed.
- *chunk*. This element is new to our DTD, compared to the original xcesAna dtd. A chunk represents a phrase which does not contain chunks of the same type. This restriction, however, cannot be imposed through the DTD. A chunk contains a sequence (one element minimally) of chunks or tokens. The chunk element gets, in addition to the global attributes, an attribute *category* with the chunk categories as values. Let me know if you want to have another name for this attribute.
- *token*. This is the most important and most content-rich element of the annotation structure. The original xcesAna dtd applies the following content model:
  - the token form as it appears in the text is the textual content of the *token* element. The *token* elements contains the attributes *base* (base form, also called lemma), *ctag* (part of speech), *msd* (morphosyntactic description). The value of the latter is a string of morphosyntactic features. Currently, the *msd* element allows for simple, unstructured textual content. This mirrors the format of the EAGLES morphosyntactic tags as a vector of feature values, as can be seen in the following examples:

**example 3:** *Afpms-* (example taken from the xcesAna documentation)

**example 4:** *sg:nom:m1:pos* (example from the Polish corpus)

## 5 The DTD

```

<!-- -->
<!-- -->
<!--          Corpus Encoding Standard -->
<!-- -->
<!--          CES -->
<!-- -->
<!--          Encoding conventions for annotated data -->
<!-- -->
<!-- -->
<!--          Modified for the LT4EL project -->
<!--          (annotation of learning objects) -->
<!--          Version LT4EL-1.0release -->
<!--          This version covers the purely linguistic -->
<!--          which is used in applicaion contexts -->
<!-- -->
<!-- -->
<!-- -->
<!-- -->
Original Date: 1996/08/05 19:07:30
Original Revision: 1.11
This is a modification of the original CES DTD designed for use with
the LT4EL learning objects. Created by Lothar Lemnitzer
and Adam Przepiorkowski.
See the accompanying documentation for changes.
All changes following version 1.0 will be documented inline.

```

This version is changed by Kiril Simov in order to facilitate the annotation. It can be transformed into the original with little effort.

I made all required attribute to be not obligatory in order to not cause problems during the annotation. The only required attribute is currently the ID which has to be given for the linguistic objects (paragraph, sentence, chunk, token). For testing, this requirement can be alleviated, but the resulting documents *\*must\** conform to this stricter version.

Also I introduce temporal attribute that will help the annotation process, but they will miss in the last version of the corpus.

I also delete all marked element and substitute them with attribute rend.

In this version the header is deleted too.

Thus, additionally, I deleted <orth> element, assuming that the immediate text within a <tok> element is the orthography of the token.

This version reflects the discussion and decision of the LT4EL TC as of 14 June 2006

1. to remove the markedX elements and place the layout information which is passed on from the baseXML documents to the linguistically annotated documents in a "rend" attribute

2. to use feature vectors as values of the attribute "msd"
- 3.a. to remove the elements orth and lex without any substitute;
- 3.b. to transform the elements ctg and msd into attributes of the element tok
4. to allow for markedTerm elements within markedTerm elements.

On the Wiki:

<http://www.let.uu.nl/lt4el/wiki/index.php/Image:LT4ELAnaProject-v3.dtd>

Version 3.1

Smaller change: it is not longer necessary to provide an id for the <LT4ELAna> root element

```
-->

<!--          Global attributes          -->

<!ENTITY % a.global '
id ID          #REQUIRED
xml:lang CDATA #IMPLIED
lang CDATA    #IMPLIED
rend CDATA    #IMPLIED'>

<!ENTITY % a.ana '%a.global;
type CDATA    #IMPLIED
wsd CDATA    #IMPLIED'
>

<!-- The following are language specific attributes that are used
temporary during the annotation of Bulgarian, Czech, etc
they will be deleted when the learning objects are
delivered -->

<!ENTITY % bg.attrs '
aa CDATA #IMPLIED
ana CDATA #IMPLIED
cat CDATA #IMPLIED
exp CDATA #IMPLIED' >

<!ENTITY % cz.attrs '' >

<!ENTITY % de.attrs '' >

<!ENTITY % temp.attrs '%bg.attrs;%cz.attrs;%de.attrs;
' >

<!-- We can imagine that each group would like to have such temporal attributes.
When we want to exclude these attributes, we have to give empty definition
for the temp.attrs ENTITY
-->

<!ELEMENT LT4ELAna (par+) >
<!ATTLIST LT4ELAna
      id          ID          #IMPLIED
      xml:lang    CDATA      #IMPLIED
      lang        CDATA      #IMPLIED
      rend        CDATA      #IMPLIED
type          CDATA          #IMPLIED
```

```

        wsd          CDATA          #IMPLIED
version CDATA #IMPLIED >

<!ELEMENT par (s | tok)+ >
<!ATTLIST par %a.ana;
name CDATA #IMPLIED >

<!ELEMENT s (chunk | tok | markedTerm | definingText)+ >
<!ATTLIST s %a.ana;
>

<!ELEMENT chunk (chunk | tok | markedTerm)+ >
<!ATTLIST chunk %a.ana;
category CDATA #IMPLIED >

<!ELEMENT tok          (#PCDATA)          >
<!ATTLIST tok %a.ana;%temp.attrs;
sp (y|n) "n"
name CDATA #IMPLIED
class CDATA          #IMPLIED
base CDATA          #IMPLIED
ctag CDATA          #IMPLIED
msd CDATA          #IMPLIED
>

<!-- Definition of the attributes:
sp - determines whether after the token there was a space or not
(it makes sense only if spaces are deleted after the tokenization;
name - the address of the tok element before the conversion from
Basic XML into LT4ELana XML. For the element par it
plays the same role
class - General classification of the tokens in the text: word, number, punctuation etc
base - The base form for the wordform (token)
ctag - Part of speech info
msd - morphosyntactic description
-->

<!-- The following elements are specific to the project
and even more specific to the WP2 tasks
the model the keyword and defition elements to be annotated -->

<!ELEMENT markedTerm (chunk | tok | markedTerm)+ >
<!ATTLIST markedTerm
%a.ana;
kw (y|n) "n"
dt (y|n) "n"
status CDATA #IMPLIED
comment CDATA #IMPLIED >

<!ELEMENT definingText (chunk | tok | markedTerm)+ >
<!ATTLIST definingText
id ID #IMPLIED
xml:lang CDATA #IMPLIED
lang CDATA #IMPLIED

```

rend	CDATA	#IMPLIED	
type	CDATA	#IMPLIED	
wsd	CDATA	#IMPLIED	
def	IDREF	#IMPLIED	
continue	CDATA	#IMPLIED	
part	CDATA	#IMPLIED	
status	CDATA	#IMPLIED	
comment	CDATA	#IMPLIED	>