The TUSNELDA annotation standard: An XML encoding standard for multilingual corpora supporting various aspects of linguistic research

Laura Kallmeyer, Andreas Wagner
SFB 441
University of Tuebingen
{lk,wagner}@sfs.nphil.uni-tuebingen.de

1. Introduction

This paper proposes a corpus encoding standard that meets the needs of linguistic research using a variety of linguistic data structures. The standard was developed in SFB 441, a research project at the University of Tuebingen.

The principal concern of SFB 441 are the empirical data structures which feed into linguistic theory building. SFB 441 consists of several projects, most of which are building corpora to empirically investigate various linguistic phenomena in various languages (e.g. modal verbs in German, forms of address and politeness in Russian). These corpora will form the components of the "Tuebingen collection of reusable, empirical, linguistic data structures (TUSNELDA)".

The TUSNELDA annotation standard aims at providing a uniform encoding scheme for all subcorpora and texts of TUSNELDA such that they can be processed with uniform standardized tools. To guarantee maximal reusability we use XML for encoding. Previous SGML standards for text encoding were provided by the Text Encoding Initiative (TEI) and the Expert Advisory Group on Language Engineering Standards (Corpus Encoding Standard, CES). The TUSNELDA standard is based on TEI and XCES (XML version of CES) but takes into account the specific needs of the SFB projects, i.e. the peculiarities of the examined languages and linguistic phenomena.

2. General structure of TUSNELDA

The overall structure of a TUSNELDA corpus is inspired by XCES. A corpus consists of a header and either one or more documents or one or more subcorpora. A document then contains a header and a text.

As in TEI and XCES, a header may have four subelements: The file description with information about the corpus itself or the texts within it, the encoding description that concerns the relation between the electronic text and its source, the profile description giving information about various non-bibliographic aspects of a text, and the revision description that provides the revision history of the file. The structure of texts is more or less as in XCES.

In the following sections, we describe the main differences between TUSNELDA and XCES.

3. Maintaining uniformity throughout TUSNELDA

In a research group like SFB 441 with different projects encoding corpora in different languages with different linguistic annotations, a uniform markup approach must be guaranteed to obtain overall comparability. Therefore, in several respects, TUSNELDA is stricter than XCES: some elements that are optional in XCES are required in TUSNELDA and for some attributes, the possible values are more restricted than in XCES.

In the header, the value of "type" is restricted to "text" or "corpus". In TEI and in XCES, "type" is intended to have these two values, but its value is defined as CDATA. The restriction avoids for example the use of "korpus" instead of "corpus".

Further, elements encoding version and revision history of the corpus which were optional in XCES are required in TUSNELDA.

In TUSNELDA, for all cases of correction and normalization (e.g. w.r.t. spelling) it is highly recommended to keep the original form. Therefore the attributes "method" of <correction> and <normalization> respectively (both part of <editorialDecl>, a subelement of the encoding description) have default values "tags" instead of "silent" as in XCES.

4. Technically motivated extensions of XCES

The element <extent>, subelement of the file description, gives the size of the electronic text. In XCES only subelements <wordCount> and <byteCount> are provided. <byteCount> contains the count of bytes in the file (text and markup) whereas <wordCount> contains the count of words in the text. This count is often used to specify the size of a corpus. One problem with the XCES <wordCount> is that it may or may not include punctuation marks. However, it is desirable to take into account both cases. Therefore, we added a new element <tokenCount> counting words and punctuation marks besides <wordCount> that leaves punctuation marks aside. Furthermore, we introduced an additional <characterCount> containing the number of characters of the text without markup. For some (agglutinating or highly inflectional) languages, the number of characters is more informative w.r.t. the size of a corpus than the number of words.

The element <segmentation> which is part of the encoding description states the principles according to which the text has been segmented into tag contents, e.g. into sentences. In XCES <segmentation> is not structured further. For corpora like TUSNELDA with a variety of tags (words, sentences, part-of-speech, syntactic chunks), information about how these annotations were obtained is very helpful for the evaluation and reusability of tools and corpora. Therefore we redefined <segmentation> as containing arbitrarily many pairs of a tag <tag> and the corresponding segmentation method <segmMethod>. All tools used for segmentation w.r.t. a certain tag should be named in its <segmMethod>.

The CES definition of the element <sp> (speech) contains a subelement <stage> for stage directions that can appear anywhere below <sp>. XCES uses <stage> as a paragraph-level element, i.e. <sp> consists of subelements <speaker>, <p> (paragraph) and <stage>. This solution however is problematic. Firstly, the grouping of a speaker and his text in <sp> that exists in CES is no longer given in XCES: The XCES <sp> allows any number of <speaker>, <p> and <stage> elements in any order. Secondly, cases where stage directions occur within a paragraph cannot be adequately annotated.

To solve these problems, we defined a new element <spokenPar> that is similar to <p> but may also contain <stage> besides phrase sequences. <sp> is then redefined containing 0 or more elements <speaker> followed by one or more elements <spokenPar> or <stage>. This allows <stage> to appear at both the paragraph and subparagraph level, and ensures that with a new speaker specification, a new element <sp> begins.

As an example consider the following:

```
<sp who="Lady Windermere">
    <speaker>Lady Windermere.</speaker>
    <spokenPar>That will do!</spokenPar></sp>
<sp><stage>Exit Parker C.</stage></sp>
<sp who="Lady Windermere">
    <spokenPar><stage>Speaking to Lord Windermere</stage>
       Arthur, if that woman comes here - I warn you -
    </spokenPar></sp>
```

5. Specific corpus-linguistic needs

Sentences can be nested, i.e. one sentence may contain another one (e.g. a quotation). As it may be interesting to examine the properties of such nested sentences in contrast to non-nested sentences, we introduced the attribute "nested" for the element <s> (sentence). The possible values of "nested" are "yes" and "no" (the latter being the default). A sentence is classified as nested if it contains another sentence. With this explicit encoding, nested and non-nested sentences can be distinguished more easily.

Each language used in a text or subcorpus is declared in a <language> element in the header. XCES defines the obligatory attribute "iso639" for <language> containing a language code from ISO 639 (e.g. "en" for English). However, the ISO standard does not cover all the languages and dialects that will be included in TUSNELDA. ISO 639-2 comprises 460 languages. Another standard, Ethnologue (Grimes 1999), comprises 6,703 languages and dialects. Therefore, we added the optional attribute "ethnologue" that allows to provide the Ethnologue language code where necessary. We kept the "iso639" attribute because of the prominence of the ISO standard.

TUSNELDA will also contain diachronic collections of texts. For such texts, knowledge about the date and the place of their creation is crucial. To explicitly capture these data, we extended the <creation> element (in the profile description), which keeps information about the origin of a text, by adding three attributes: "place", to specify the place of creation; "earliest" and "latest", to delimit the period of time during which the text was created (often it is not possible to determine an exact date of creation for historical texts).

As mentioned in Section 3, if corrections and normalizations are applied, the original form should be preserved. This policy may cause problems for historical texts. Here it might be the case that some portion is not uniquely recoverable because the original document is damaged. The same problem arises for transcriptions of speech recordings which contain noise. For such cases, we introduced the new sub-paragraph element <unclear>, which is not defined in XCES but in the TEI guidelines. <unclear> is intended to contain reconstructions of such damaged portions for which it is not possible to provide the original.

Some projects of the SFB 441 investigate specific linguistic phenomena, and in order to do so, they collect corpora to search for a certain class of linguistic elements. One project, for example, is interested in deictic expressions, another project in temporal adverbial modifiers. For the automatic retrieval of these specific elements, exhaustive Part-of-Speech tagging is not necessary and in some cases even not sufficient. Instead, we introduced a new phrase-level element <marked> with the obligatory attribute "type". With this element one can tag e.g. temporal adverbials as <marked type="adv-tmp">. <marked> provides a flexible means to identify exactly those elements that are relevant for the intended application of the acquired corpora.

6. Conclusion

We have developed a corpus annotation standard by adapting the XCES to the requirements of our corpus collection TUSNELDA. This standard takes into account the needs of the various linguistic research projects of SFB 441 (for example by explicitly encoding crucial information) while ensuring a standardized markup that allows uniform processing of all subcorpora.

It should be pointed out that, despite the above-mentioned modifications of XCES, most parts of XCES could be adopted for TUSNELDA without changes. This shows that, although XCES was developed for language engineering tasks, it is in essence suitable for theoretical linguistic research as well.

Bibliography

Expert Advisory Group on Language Engineering Standards (EAGLES)
Corpus Encoding Standard - Document CES 1. Version 1.5. 27 January 1999.
http://www.cs.vassar.edu/CES/

Expert Advisory Group on Language Engineering Standards (EAGLES)
XCES Corpus Encoding Standard for XML. XML version of the CES DTDs.
Document XCES 0.2. 10 February 2000.
http://www.cs.vassar.edu/XCES/

Grimes, Barbara F. (ed.)
Ethnologue: Languages of the World, Thirteenth Edition.
SIL Publications. 1996.

Kallmeyer, Laura and Andreas Wagner
Guidelines for the TUSNELDA Corpus Annotation Standard.
http://www.sfb441.uni-tuebingen.de/c1/tusnelda_guidelines.html
To appear.

Sperberg-McQueen, C.M., Burnard, L. (eds.)
Guidelines for Electronic Text Encoding and Interchange.
Text Encoding Initiative, Chicago and Oxford. 1994.
http://etext.virginia.edu/TEI.html