TEI P5 as a Text Encoding Standard for Multilevel Corpus Annotation

Bański, Piotr
pkbanski@uw.edu.pl
University of Warsaw

Adam Przepiórkowski
adamp@ipipan.waw.pl
Institute of Computer Science Polish Academy of Sciences

The need for text encoding standards for language resources (LRs) is widely acknowledged: within the International Standards Organization (ISO) Technical Committee 37 / Subcommittee 4 (TC 37 / SC 4), work in this area has been going on since the early 2000s, and working groups devoted to this issue have been set up in two current pan-European projects, CLARIN (`http://www.clarin.eu/`) and FLaReNet (`http://www.flarenet.eu/`). It is obvious that standards are necessary for the interoperability of tools and for the facilitation of data exchange between projects, but they are also needed within projects, especially where multiple partners and multiple levels of linguistic data are involved.

One such project is the National Corpus of Polish (Pol. *Narodowy Korpus Języka Polskiego*; NKJP; `http://nkjp.pl/`; Przepiórkowski *et al.* 2008, 2009) involving 4 Polish institutions and carried out in 2008–2010. The project aims at the creation of a 1-billion-word automatically annotated corpus of Polish, with a 1-million-word subcorpus annotated manually. The following levels of linguistic annotation are distinguished in the project: 1) segmentation into sentences, 2) segmentation into fine-grained word-level tokens, 3) morphosyntactic analysis, 4) coarse-grained syntactic words (e.g., analytical forms, constructions involving bound words, etc.), 5) named entities, 6) syntactic groups, 7) word senses (for a limited number of ambiguous lexemes).

Any standards adopted for these levels should allow for stand-off annotation, as is now common practice and as is virtually indispensable in the case of many levels of annotation, possibly involving conflicting hierarchies.

Two additional, non-linguistic levels of annotation required for each document are text structure (e.g., division into chapters, sections and paragraphs, appropriate marking of front matter, etc.) and metadata. The standard adopted for these levels should be sufficiently flexible to allow for representing diverse types of texts, including books, articles, blogs and transcripts of spoken data.

NKJP is committed to following current standards and best practices in corpus development and text encoding. However, because of the current proliferation of official, *de facto* and purported standards, it is far from clear what standards a new corpus project should adopt. The aim of this paper is to attempt to answer this question.

1. Standards and best practices

The three text encoding standards and best practices listed in a recent CLARIN short guide (CLARIN:STE, 2009)[1] are: standards developed within ISO TC 37 SC 4, the Text Encoding Initiative (TEI; Burnard and Bauman 2008) guidelines and the XML version of the Corpus Encoding Standard (XCES; Ide *et al.* 2000). Apart from these, there are other *de facto* standards and best practices, e.g., TIGER-XML (Mengel and Lezius, 2000) for the encoding of syntactic information, or the more general PAULA (Dipper, 2005) encoding schema used in various projects in Germany.

1.1. XCES

The original version of XCES inherits from TEI an exhaustive approach to metadata representation. It makes specific recommendations for the representation of morphosyntactic information and for the alignment of parallel corpora. In early the 2000s, it was probably the most popular corpus encoding standard.

Currently, the claim of XCES to being such a standard is much weaker. A new — more abstract — version of XCES was introduced around 2003, where concrete morphosyntactic schema was replaced by a general feature

structure mechanism, different from the ISO Feature Structure Representation (FSR) standard (ISO 24610-1). In our view, this is a step back, as adopting a more abstract representation requires more work on the part of corpus developers. Moreover, XCES has no specific recommendations for other levels of linguistic knowledge, and no mechanisms for representing discontinuity and alternatives, all of which need to be represented in NKJP. Taking also into account the lack of documentation and the potential confusion concerning its versioning,[2] XCES turns out to be unsuitable for the purposes of NKJP.

1.2. ISO TC37 SC 4

There is a family of ISO standards developed by ISO TC 37 SC 4 for modelling and representing different types of linguistic information. The two published standards concern the representation of feature structures (ISO 24610-1) and the encoding of dictionaries (ISO 24613). Other proposed standards are at varying levels of maturity and abstractness. While eventually these standards may reach stability and specificity required by practical applications, this is currently not the case.[3]

1.3. TIGER-XML and PAULA

TIGER-XML and a schema which may be consider as its generalisation, PAULA, are specific, relatively well-documented and widely employed best practices for describing linguistic objects occurring in texts (so-called "markables") and relations between them (in the case of TIGER-XML, the constituency relation). They do not contain specifications for metadata or structural annotation.

2. TEI P5

For metadata and structural annotation levels there is no real alternative to TEI. Moreover, TEI P5 implements the FSR standard ISO 24610-1, which can be used for the representation of any linguistic content, along the lines of XCES (although the feature structure representations used in XCES do not comply with this standard), PAULA and the proposed ISO standard, Linguistic Annotation Framework (ISO 24612). TEI P5 is stable, has rich documentation and an active user base, and for these reasons alone it should be preferred to XCES and (the current versions of) the ISO standards. Moreover, any TIGER-XML and PAULA annotation may be expressed in TEI in an isomorphic way, thanks to the linking mechanisms of TEI P5.

However, TEI is a very rich toolbox, proposing multitudinous mechanisms for representing multifarious aspects of text encoding, and this richness, as well as the sheer size of TEI P5 documentation (1350–1400 pages), are often perceived by corpus developers as prohibitive. For this reason, within NKJP, a specific set of recommendations for particular levels of annotation has been developed, aiming at achieving a maximal compatibility (understood as the easiness to translate between formats) with other proposed and *de facto* standards.

For example, TEI P5 offers, among others, the following ways to represent syntactic constituency:

- XML tree, built with elements such as <s>(entence), <phr>(ase), <cl>(ause) and <w>(ord), may directly mirror constituency tree;
- all information, including constituency, may be encoded as a feature structure (Witt *et al.*, 2009);
- each syntactic group is a <seg>(ment) of type group, containing a feature structure description and <ptr> pointers to other constituents, defined in the same file (for non-terminal syntactic groups) or in a stand-off way elsewhere (for terminal words).

While the first of these representations is the most direct, and the second most general, it is the third representation that directly mirrors TIGER-XML, PAULA and SynAF, and for this reason, it has been adopted in NKJP.

References

Bel, N., Beskow, J., Boves, L., Budin, G., Calzolari, N., Choukri, K., Hinrichs, E., Krauwer, S., Lemnitzer, L., Piperidis, S., Przepiórkowski, A., Romary, L., Schiel, F., Schmidt, H., Uszkoreit, H., and Wittenburg, P. (2009). *Standardisation action plan for Clarin. State: Proposal to CLARIN Community*.

Burnard, L. and Bauman, S. (ed.) (2008). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. http://www.tei-c.org/Guidelines/P5/.

CLARIN:STE (ed.) (2009). *Standards for text encoding: A CLARIN shortguide*. http://www.clarin.eu/documents.

Dipper, S. (2005). 'Stand-off representation and exploitation of multi-level linguistic annotation'. *Proceedings of Berliner XML Tage 2005 (BXML 2005)*. Berlin, pp. 39-50.

Ide, N., Bonhomme, P., and Romary, L. (2000). 'XCES: An XML-based standard for linguistic corpora'. *LREC*. **2000**: 825-830.

LREC (2000). 'Proceedings of the Second International Conference on Language Resources and Evaluation, LREC 2000'. *ELRA*. Athens.

Mengel, A. and Lezius, W. (2000). 'An XML-based encoding format for syntactically annotated corpora'. *LREC*. **2000**: 121-126.

Przepiórkowski, A., Górski, R. L., Lewandowska-Tomaszczyk, B., and Łaziński, M. (2008). 'Towards the National Corpus of Polish'. *Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC 2008*. Marrakech: ELRA.

Przepiórkowski, A., Górski, R. L., Łaziński, M., and Pęzik, P. (2009). 'Recent developments in the National Corpus of Polish'. *Proceedings of Slovko 2009: Fifth International Conference on NLP, Corpus Linguistics, Corpus Based Grammar Research, 25–27 November 2009, Smolenice/Bratislava, Slovakia*. Levická, J and Garabík, R. (ed.). Brno. Tribun.

Witt, A., Rehm, G., Hinrichs, E., Lehmberg, T., and Stemann, J. (2009). 'SusTEInability of linguistic resources through feature structures'. *Literary and Linguistic Computing*. **24(3)**: 363-372.

Notes

1. See also Bel *et al.* 2009.
2. Two different sets of schemata have co-existed on XCES WWW pages since 2003, one given as DTD, another as XML Schema, without any clear indication that they specify different structures.
3. A tendency may be observed of increasing abstractness and generality of proposed standards, esp., SynAF (ISO 24615) and LAF (ISO 24612)". This leads to their greater formal elegance, at the cost of their actual usefulness.