TEI P5 as an XML Standard for Treebank Encoding*

Adam Przepiórkowski

Institute of Computer Science, Polish Academy of Sciences
and Institute of Informatics, University of Warsaw
E-mail: adamp@ipipan.waw.pl

Abstract

The aim of the paper is to show that a subset of Text Encoding Initiative Guidelines is a reasonable choice as a standard for stand-off XML encoding of syntactically annotated corpora. The proposed TEI schema — actually employed in the National Corpus of Polish — is compared to other such candidate standards, including TIGER-XML, SynAF and PAULA.

1 Introduction

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.

Given the existence of a number of proposed and *de facto* standards for various levels of linguistic annotation, the starting point of the design of a specific schema to be used in a particular project should be careful examination of those standards: in the interest of interoperability, creation of new schemata should be avoided. But, if a number of standards are applicable, which one should be chosen and on what grounds? And if constructing a schema for a particular linguistic level in a particular project should start with an overview and comparison of existing

*The work reported here, carried out within the National Corpus of Polish project funded in 2007–2010 by the research and development grant R17 003 03 from the Polish Ministry of Science and Higher Education, has its origins in a series of papers co-authored by Piotr Bański, dealing with the encoding of lower levels of linguistic information and with the metadata; cf. Bański and Przepiórkowski 2009 and Przepiórkowski and Bański 2009a,b. The paper also profited from insightful comments by TLT8 reviewers.

standards — a time consuming task which may be hard to justify within the limits of a project budget — would it not be easier and cheaper to construct one's own focused schema from scratch?

The aim of this paper is to examine the most popular standards and to make specific recommendations concerning the encoding of syntactic information. We look at four standards: 1) the very specific and commonly used TIGER-XML schema, 2) SynAF, a more general model derived from TIGER-XML, currently under development as an ISO standard, 3) PAULA, a schema for the representation of various linguistic levels, and 4) version P5 of the Text Encoding Initiative (TEI) Guidelines, proposing multitudinous mechanisms for representing multifarious aspects of text encoding. Perhaps because of this richness, TEI is the least obvious candidate for treebank encoding, but it is the one that we would like to argue for here.

2 Standards and Best Practices for Treebank Encoding

In this section we briefly examine three probably most often cited standards and best practices for treebank encoding: TIGER-XML, SynAF (and related ISO proposed standards) and PAULA. A TEI P5 schema for the annotation of syntactic information is presented in § 3, and a discussion of relative merits of these standards is given in § 4.

2.1 TIGER-XML

TIGER-XML (Mengel and Lezius 2000) is a *de facto* standard for XML annotation of treebanks. It is well documented[1] and exemplified, it has been adopted in various projects, and it was the starting point for the SynAF proposed standard.

In TIGER-XML, each sentence is represented as a `<graph>` consisting of `<terminals>` and `<nonterminals>`. The `<terminals>` element is a list of `<t>`erminals, with orthographic, morphosyntactic and other information represented in attributes. Morphosyntactic attributes and their possible values may be defined in corpus `<head>`er.

Similarly, `<nonterminals>` is a list of `<nt>` syntactic nodes. Within each node, `<edge>`s link the node to its immediate constituents (`<t>`s or `<nt>`s). Additional secondary edges (`<secedge>` elements within `<nt>`) may be used to represent co-reference or other non-constituency information.

There is a treebank search engine serving TIGER-XML corpora, TIGERSearch (Lezius 2002, König *et al.* 2003), and converters from TIGER-XML to other formats, including the PAULA format used by ANNIS2 (`http://www.sfb632.uni-potsdam.de/d1/annis/`) and the Poliqarp (Janus and Przepiórkowski 2007) format.

[1] `http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/doc/html/TigerXML.html`.

2.2 SynAF and Related Standards

Two proposed ISO standards immediately relevant for syntactic annotation are Syntactic Annotation Format (SynAF; ISO 24615) and Morphological Annotation Format (MAF; ISO 24611). According to `http://www.iso.org/` both were at the DIS stage of ISO standard development[2] at the time of writing this paper, although only the CD versions were available free of charge at `http://www.tc37sc4.org/`. It is these CD versions, ISO:24615 2009 and ISO:24611 2005 (see the bibliography), that we refer to here.

While TIGER-XML provides a specific schema, "SynAF is dealing with the description of a *metamodel* for syntactic annotation" (ISO:24615 2009, p. 6; emphasis ours), where syntactic annotation includes both constituency and dependency marking. The SynAF metamodel is described within a page and a half (ISO:24615 2009, pp. 12–14) as a straightforward generalisation of TIGER-XML.[3] The three main classes are: *T_Node* (for terminal nodes), *NT_Node* (for non-terminal nodes) and *Edge* (for dependency edges between nodes). Both kinds of nodes are defined over spans of text.

According to SynAF, syntactic annotation is applied to MAF-annotated input. MAF (ISO:24611 2005) is a much more specific and mature standard than SynAF, providing various examples of the encoding of morphosyntactic information. The two main elements it provides are `<token>` and `<wordForm>`. Tokens may identify spans in an external file (stand-off annotation), or may be marked in the original document (embedded annotation). Token attributes such as `@form` and `@transcription` may be used to abstract over the sequence of characters marked as a `<token>`. The `@join` attribute may specify whether there is whitespace to the left or right of the token. Two or more `<token>`s may overlap, e.g., in the case of abbreviations, where the final dot may belong to the abbreviation `<token>` and also constitute a separate `<token>`.

Word forms are defined over tokens. In the default case, one `<wordForm>` corresponds to one `<token>` and adds information about the lemma, the morphosyntactic analysis, etc. This linguistic information may be encoded as attributes of `<wordForm>`s or as feature structures (ISO:24610-1 2005) within `<wordForm>`s.[4] Word forms may also be empty (as for *pro* or PRO in Chomsky's

[2]There are six stages of development of any ISO standard: 1) initial proposal of a new work item, 2) preparation of a Working Draft (WD), 3) production and acceptance of the Committee Draft (CD), 4) production and acceptance of the Draft International Standard (DIS), to be distributed to ISO member bodies for commenting and voting, 5) approval of the Final Draft International Standard (FDIS), which has to pass the final vote, and 6) the publication of the International Standard (IS).

[3]Unfortunately, most of ISO:24615 2009 is devoted to an annex containing a preliminary list of syntactic data categories, even though, in our opinion, the contentious issue of linguistic data categories should be kept separate from the relatively straightforward matter of defining a metamodel or specific XML encoding of constituency and dependency relations. The other annex contains "testsuites", i.e., examples of MAF and SynAF annotation, which contain errors (they are not well-formed XML) and need "to be considerably revised" (p. 63). In summary, the practical usefulness of the version of SynAF referred to here is still rather limited.

[4]MAF contains also recommendations on tagset specification and on the handling of various

Principles and Parameters), or they may consist of a number of possibly discontinuous tokens. One <token> may also give rise to a number of <wordForm>s (as, possibly, in the case of German *am*, French *auquel*, Italian *damelo*, English *wanna* or Polish *nań*, if they are analysed as single <token>s). Moreover, <wordForm>s may contain other <wordForm>s, e.g., for the purpose of representing various multi-word units, so the domain of applicability of MAF overlaps somewhat with the domain of SynAF. Word forms may also refer to an external lexicon for their definitions.

Another proposed ISO standard that should also be mentioned here is Linguistic Annotation Framework (LAF; ISO:24612 2008), which defines a generic graph-based pivot format, called GrAF, designed to facilitate comparison and exchange of data in various annotation formats.

2.3 PAULA

PAULA (Ger. *Potsdamer AUstauschformat für Linguistische Annotation*; Dipper 2005), a LAF-inspired format developed within the SFB 632 project in Potsdam and Berlin, is an example of a family of general encoding standards for the annotation of multi-modal data.[5]

In the PAULA data model there are objects ("markables"), various types of relations between them, and features of objects. Markables may be simple spans of text (<mark>) or abstract <struct>ures bearing <rel>ations to other markables. For example, a syntactic constituent with 3 immediate daughters (one word and two syntactic constituents) may be represented as follows:[6]

```
<struct id="syn2"> <!-- PAULA -->
 <rel id="rel3" type="head"    xlink:href="tok.xml#t6"/>
 <rel id="rel4" type="nonhead" xlink:href="#syn20"/>
 <rel id="rel6" type="nonhead" xlink:href="#syn21"/>
</struct>
```

This representation closely corresponds to the following representation in TIGER-XML, though PAULA's <rel> is a generalisation of TIGER-XML's <edge> and may be used for the representation of various types of relations.

```
<nt id="nt2"> <!-- TIGER-XML -->
 <edge label="head"    idref="#t6"/>
 <edge label="nonhead" idref="#nt20"/>
 <edge label="nonhead" idref="#nt21"/>
</nt>
```

Additionally, markables are associated with feature values via a PAULA-specific encoding of feature structures.

---

kinds of ambiguities, including structural ambiguities encoded as finite state automata.

[5] See Dipper *et al.* 2006 for references to other such largely graph-based encodings.

[6] This is a modification of an example from Dipper 2005.

3 TEI P5

The Text Encoding Initiative "was established in 1987 to develop, maintain, and promulgate hardware- and software-independent methods for encoding humanities data in electronic form" (`http://www.tei-c.org/`). It is a *de facto*, constantly maintained XML standard for encoding and documenting textual data, with an active community, detailed guidelines (Burnard and Bauman 2008) and supporting tools. Its recommendations for the encoding of linguistic information are limited, but it includes the ISO FSR standard for representing feature structures, which can be used to encode various kinds of information (cf., e.g., Witt *et al.* 2009).

There are some TEI-encoded morphosyntactically annotated corpora, but the impact of the current P5 version of TEI Guidelines, released in November 2007, has been rather limited so far. Probably the main reason for this state of affairs is the richness and versatility of TEI. Ideas useful for linguistically annotated corpora are scattered over the 1350-odd pages of the Guidelines, and usually there is more than one way of representing any given annotation, so designing a coherent and constrained TEI-conformant schema for linguistic corpora is a daunting task.

One such schema, indirectly based on an earlier version of TEI Guidelines, is XCES (Ide *et al.* 2000), an XML-ised version of the TEI-based Corpus Encoding Standard (CES; Ide and Priest-Dorman 1995, Ide 1998) specified in SGML. XCES DTD schemata specify the representation of metadata, primary data, morphosyntactic annotation and — for parallel corpora — alignment. There are general feature structure mechanisms for the representation of other levels of information, but it is the specificity at the morphosyntactic and alignment levels that had a large influence on the relative success of that version of XCES. Around 2003 a new version of XCES was introduced — given as XML Schema specifications — that was a step back in this respect, as it lacks specific recommendations for any linguistic levels, resorting instead to general feature structure mechanisms; only minor technical modifications have been made to these schemata since their introduction. Other reasons why XCES does not currently meet the expectations of corpus developers are: 1) lack of documentation; `http://www.xces.org/` refers to old CES documentation as "supporting general encoding practices for linguistic corpora and tag usage" and "largely relevant to the XCES instantiation", although the CES documentation is hardly applicable to the second version of XCES, 2) feature structure mechanisms different from the established feature structure representation ISO standard (ISO:24610-1 2005), 3) lack of mechanisms for the encoding of discontinuity 4) or alternatives, and 5) the potential for confusion regarding the version of the standard (in particular, for many years DTD and XML Schema specifications co-existed on XCES web pages, without any clear information that they specify different representations).

In Przepiórkowski and Bański 2009b, we propose a representation of the primary data, text structure, text headers and corpus headers conformant with TEI P5. In this section we describe a possible TEI P5 encoding of syntactic information

designed to maximise compatibility with other proposed standards.[7]

3.1 Morphosyntax and Other Assumptions

Following the common standard practice, we assume that each linguistic level is represented in its own file, referring to lower layers, down to the primary text, i.e., we assume the stand-off approach to annotation. More precisely, each corpus text is represented as a collection of files containing various annotation layers of the text, and each layer has the following minimal general structure (where `corpus_header.xml` contains the unique corpus header, and `header.xml` is the header of the particular text):

```
<teiCorpus
 xmlns:xi="http://www.w3.org/2001/XInclude"
 xmlns="http://www.tei-c.org/ns/1.0">
 <xi:include href="corpus_header.xml"/>
 <TEI>
  <xi:include href="header.xml"/>
  <text>
   <body><!-- text to annotate --></body>
  </text>
 </TEI>
</teiCorpus>
```

Depending on the type of text (written or spoken), `<body>` contains a list of `<p>`aragraphs (or possibly paragraph-length anonymous blocks, `<ab>`) or `<u>`tterance turns, further split into `<s>`entences, e.g.:

```
<body>
 <p xml:id="segm_p1">
  <s xml:id="segm_s1">...</s> <!-- more sentences here -->
 </p> <!-- more paragraphs here -->
</body>
```

We assume that this structure, down to the sentence level, is preserved and made parallel at all annotation layers. The correspondence between different layers is expressed with the TEI attribute `@corresp`. For example, assuming that the above code is a fragment of the segmentation layer of a text (by assumption represented in `ann_segmentation.xml`), defining word-level tokens, the next layer, morphosyntax, may have the following parallel structure:

```
<body>
 <p xml:id="morph_p1" corresp="ann_segmentation.xml#segm_p1">
  <s xml:id="morph_s1" corresp="ann_segmentation.xml#segm_s1">...</s>
 </p>
</body>
```

Whatever other annotation layers are present in the corpus, we assume the existence of a morphosyntactic layer (by assumption encoded in

[7] The overall picture, encompassing all linguistic levels assumed in the National Corpus of Polish, is presented in Przepiórkowski and Bański 2009a.

ann_morphosyntax.xml). Each sentence at this layer is a sequence of <seg> elements implicitly (via a specification in the schema) marked as @type="token", and each <seg> contains a feature structure specification of various morphosyntactic information about the segment, e.g.:

```
<s xml:id="morph_s1" corresp="ann_segmentation.xml#segm_s1">
 <seg xml:id="morph_seg1"><fs>...</fs></seg>
 <seg xml:id="morph_seg2"><fs>...</fs></seg>
<!-- more segments here -->
</s>
```

TEI P5 contains wholesale the ISO standard for feature structure representation (ISO:24610-1 2005). In the interest of brevity and readability, we will not fully specify the XML encoding of feature structures, signalled above as <fs>...</fs>, but rather represent feature structures in a way common in linguistic theories such as HPSG and LFG. For example, the Polish segment *komputerem*, analysed as a singular instrumental inanimate-masculine form of the noun KOMPUTER, may have the following feature structure representation:

$$\begin{bmatrix} \textit{morph} & \\ \text{ORTH} & \text{komputerem} \\ \text{BASE} & \text{komputer} \\ \text{CTAG} & \text{subst} \\ \text{MSD} & \text{sg:inst:m3} \end{bmatrix}$$

Note that the names of features ORTH, BASE, CTAG and MSD are taken from (X)CES. Of course, other kinds of information may be represented here as well, including the whole list of possible interpretations (not just the one interpretation selected in the context), information about person or tool responsible for disambiguation, etc.

3.2 Representing Constituency

At the syntactic level (by assumption, in ann_syntax.xml), each <s>entence is a sequence of <seg> elements implicitly marked as @type="group". More generally, for reasons of uniformity, we propose to use the <seg> element with different values of @type for different kinds of linguistic units (see § 3.4 below), but more specific TEI elements could be used here instead, e.g., <w> for words, <phr> for syntactic phrases and <cl> for clauses.

Just like word segments, syntactic groups also contain feature structure descriptions: in the simplest case, such a description may consist of a single attribute-value pair naming the node (e.g., [NAME NP]), but it could also be an LFG f-structure or a full-fledged HPSG feature structure. Apart from a feature structure specification of the node, such a syntactic group element may contain any number of <ptr> elements pointing at the immediate constituents of the group, e.g.:

```
<seg xml:id="group2">
 <fs>...</fs>
 <ptr xml:id="ptr3" type="head" target="ann_morphosyntax.xml#seg6"/>
```

```
 <ptr xml:id="ptr4" type="nonhead" target="#group20"/>
 <ptr xml:id="ptr6" type="nonhead" target="#group21"/>
</seg>
```

Note that immediate constituents may be words specified at a different layer or other syntactic groups of the same layer. Any <ptr> element may specify the type of the constituency relation, e.g., head or nonhead, and each has an @xml:id, so that the relation may be referred to in the feature structure description of the node.

Note also that this schema allows for discontinuous constituents, as <ptr> elements within one <seg> do not have to point at neighbouring constituents. This freedom, combined with the representations for dependencies outlined in the following subsection, makes it possible to encode various linguistic analyses of possible conflicts between phrase structure and dependency, e.g., involving extraposed material or crossing dependencies.

3.3 Representing Dependency

It is equally straightforward to represent dependency relations in TEI P5: instead of the one way <ptr> pointer used for immediate constituency, the <link> element may be used to relate two syntactic nodes (words or groups). According to TEI Guidelines, <link> may be used to represent symmetrical (bidirectional) or asymmetrical (unidirectional) relations; here, by convention, <link> represents asymmetrical edges in the dependency graph, whose end vertices are specified in the value of the attribute @targets.[8]

```
<seg xml:id="group43"><fs>...</fs></seg>
<link xml:id="link17" type="subject"
 targets="ann_morphosyntax.xml#seg78 #group43"/>
```

[8]The constraint that there be exactly two references within the values of @targets may be specified in RelaxNG by constraining the TEI data model (http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-link.html) from:

```
<rng:attribute name="targets">
 <rng:list>
  <rng:ref name="data.pointer"/>
  <rng:oneOrMore>
   <rng:ref name="data.pointer"/>
  </rng:oneOrMore>
 </rng:list>
</rng:attribute>
```

to:

```
<rng:attribute name="targets">
 <rng:list>
  <rng:ref name="data.pointer"/>
  <rng:ref name="data.pointer"/>
 </rng:list>
</rng:attribute>
```

Again, the value of `@type` specifies the kind of dependency. According to the example above, the dependency of type `subject` holds between a word (defined in `ann_morphosyntax`) and a syntactic group (defined elsewhere in the same file). In a "pure" dependency treebank, where dependencies are strictly between words, `<seg>` elements may be completely absent in this layer.

3.4 Case Study: National Corpus of Polish

It is not a prerequisite of the scheme proposed here that there be exactly two layers of grammatical representation of each text, `ann_morphosyntax.xml` and `ann_syntax`, and neither are fully rooted syntactic representations necessarily assumed here. Rather, there may be various layers of various granularity.

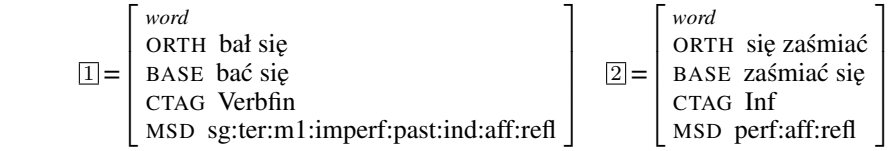
In the fully TEI P5-encoded National Corpus of Polish (Pol. *Narodowy Korpus Języka Polskiego*; NKJP; `http://nkjp.pl/`; Przepiórkowski *et al.* 2008, 2009), each text has the following linguistic layers: fine-grained word-level segmentation (including some segmentation ambiguities), morphosyntax (referring to disambiguated segmentation), coarse-grained syntactic words (e.g., for analytical tense forms consisting of multiple segments; referring to morphosyntax), named entities (referring to syntactic words) and syntactic groups (also referring to syntactic words).[9] The last layer assumes partial syntactic analysis, i.e., the annotation of nominal and other phrases, without the requirement that each word in the sentence must be contained in some syntactic group. All these layers are encoded as outlined in § 3.2, with different `@type`s of `<seg>` elements and different types of feature structure representations associated with `<seg>`s.

For example, at the syntactic words layer, `ann_words.xml`, each sentence consists of `<seg>` elements of `@type="word"` (vs. `@type="token"` for segmentation and morphosyntax). In the default case, a `<seg>` at this layer will be co-extensive with a `<seg>` at the lower (morphosyntax) layer, but it may also correspond to a possibly discontinuous list of morphosyntactic `<seg>`ments. Two different syntactic words may also overlap, as in *Bał się zaśmiać* '(He) feared (to) laugh', where for two inherently reflexive verbs, BAĆ SIĘ 'fear' and ZAŚMIAĆ SIĘ 'laugh', one occurrence of the reflexive marker *się* suffices.[10] This situation is exemplified below:

```
<seg xml:id="word13">
 <fs>1</fs> <!-- (see below) -->
 <ptr target="ann_morphosyntax.xml#seg17"/> <!-- bał -->
 <ptr target="ann_morphosyntax.xml#seg18"/> <!-- się -->
</seg>
<seg xml:id="word14">
 <fs>2</fs> <!-- (see below) -->
 <ptr target="ann_morphosyntax.xml#seg18"/> <!-- się -->
 <ptr target="ann_morphosyntax.xml#seg19"/> <!-- zaśmiać -->
</seg>
```

[9] We ignore here another layer present in NKJP, that of word senses.

[10] On the haplology of the reflexive marker in Polish, see Kupść 1999.



4 Discussion

The schema proposed above is not supposed to be novel; on the contrary, it has been designed to be as simple as possible and to be maximally compatible with other proposed standards for the encoding of grammatical information, but also with the aim of avoiding the potential problems of the other proposals.

Just like SynAF and PAULA, the schema is a straightforward extension of TIGER-XML: `<seg type="token">` (or `<seg type="word">`) elements directly correspond to TIGER's `<t>` and SynAF's *T_Node*, and `<seg type="group">` — to TIGER's `<nt>` and SynAF's *NT_Node*. Both kinds of `<seg>` elements correspond to PAULA's `<struct>`.

The schema maintains the distinction between two kinds of relations between syntactic nodes: immediate constituency, represented by `<ptr>`, and other — especially dependency — relations, represented by `<link>`, although in principle all kinds of relations could be represented via `<link>` elements, just as `<rel>` elements of different `@type`s represent different relations in PAULA. Compared to other standards, `<ptr>` corresponds to TIGER's `<edge>` and seems to have no analogue in SynAF, where apparently constituency is represented implicitly by the extent of the span of particular constituents.[11] On the other hand, `<link>` directly corresponds to SynAF's *Edge*[12] and is a generalisation of TIGER's `<secedge>`.

Moreover, various types of relations between `<token>`s and `<wordForm>`s, as defined in MAF, may be represented as illustrated in § 3.4. The current schema is also compatible with XCES, to the extent that XCES is originally TEI-based and given that the morphosyntactic representation outlined above uses XCES-inspired feature names.

We claim that the current schema inherits all advantages of various proposed standards, but improves on each of them. First of all, where TIGER-XML and MAF assume that different logical layers are present in the same file (words and syntactic groups in TIGER-XML, tokens and word forms in MAF), the schema proposed here assumes the stand-off philosophy of separating different layers of linguistic annotation.[13] Second, unlike TIGER-XML, which does not employ any feature structure representation, and unlike XCES and PAULA, which use non-standard feature structure representations, the schema proposed above complies

[11] The version of SynAF referred to in this paper is vague on this and various other specific issues.

[12] Although, curiously, in the Annex B of ISO:24615 2009, `<edge>` elements specify only one end of the edge.

[13] But such merging of annotation layers is still possible: `<seg>` elements of different `@type`s may occur in the same XML file.

fully with the ISO standard on feature structure representation. Moreover, unlike SynAF, whose current version seems to be an early draft, the schema is an application of TEI P5, a well-established and constantly maintained standard with stable guidelines and a large supporting community. In fact, since the schema is embedded in TEI, it is almost infinitely extendable and may draw from a variety of TEI solutions for various aspects of text representation.

5 Conclusion

One disadvantage of the Text Encoding Initiative P5 standard is that the documentation is huge and the task of distilling a manageable corpus schema is daunting. We have performed this task and reported the results in this and related papers (Przepiórkowski and Bański 2009a,b). Moreover, we looked at other proposed corpus encoding standards and concluded that whatever they offer may already be found in TEI, which has the advantage of being a very mature and at the same time actively supported standard. Nevertheless, whenever TEI provided alternative solutions, we chose mechanisms compatible with other proposed standards for treebank encoding, thus attaining a TEI schema maximally isomorphic with TIGER-XML, SynAF and PAULA. We hope that this work will serve as a starting point for the design of other TEI P5 corpus encoding schemata.

References

Bański, P. and Przepiórkowski, A. (2009). Stand-off TEI annotation: the case of the National Corpus of Polish. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III) at ACL-IJCNLP 2009*, pages 64–67, Singapore.

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

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

Dipper, S., Hinrichs, E., Schmidt, T., Wagner, A., and Witt, A. (2006). Sustainability of linguistic resources. In E. Hinrichs, N. Ide, M. Palmer, and J. Pustejovsky, editors, *Proceedings of the LREC 2006 Workshop on Merging and Layering Linguistic Information*, pages 14–18, Genoa. ELRA.

Ide, N. (1998). Corpus Encoding Standard: SGML guidelines for encoding linguistic corpora. In *Proceedings of the First International Conference on Language Resources and Evaluation, LREC 1998*, pages 463–470, Granada. ELRA.

Ide, N. and Priest-Dorman, G. (1995). Corpus encoding standard. `http://www.cs.vassar.edu/CES/`, accessed on 2009-08-22.

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

ISO:24610-1 (2005). Language resource management – feature structures – part 1: Feature structure representation. ISO/DIS 24610-1, 2005-10-20.

ISO:24611 (2005). Language resource management – Morpho-syntactic annotation framework (MAF). ISO/CD 24611, ISO TC 37/SC 4 document N 225 of 2005-10-15.

ISO:24612 (2008). Language resource management – Linguistic annotation framework. ISO/WD 2461[2], ISO TC 37/SC 4 document N 463 rev00 of 2008-05-12.

ISO:24615 (2009). Language resource management – Syntactic annotation framework (SynAF). ISO/CD 24615, ISO TC 37/SC 4 document N 421 of 2009-01-30.

Janus, D. and Przepiórkowski, A. (2007). Poliqarp: An open source corpus indexer and search engine with syntactic extensions. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 85–88, Prague.

Kupść, A. (1999). Haplology of the Polish reflexive marker. In R. D. Borsley and A. Przepiórkowski, editors, *Slavic in Head-Driven Phrase Structure Grammar*, pages 91–124. CSLI Publications, Stanford, CA.

König, E., Lezius, W., and Voormann, H. (2003). *TIGERSearch 2.1: User's Manual*. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Lezius, W. (2002). TIGERSearch — ein Suchwerkzeug für Baumbanken. In S. Busemann, editor, *Proceedings der 6. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2002)*, Saarbrücken.

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

Mengel, A. and Lezius, W. (2000). An XML-based encoding format for syntactically annotated corpora. In LREC (2000), pages 121–126.

Przepiórkowski, A. and Bański, P. (2009a). Which XML standards for multilevel corpus annotation? In *Proceedings of the 4th Language & Technology Conference*, Poznań, Poland. Forthcoming.

Przepiórkowski, A. and Bański, P. (2009b). XML text interchange format in the National Corpus of Polish. In S. Goźdź-Roszkowski, editor, *The proceedings of Practical Applications in Language and Computers PALC 2009*, Frankfurt am Main. Peter Lang. Forthcoming.

Przepiórkowski, A., Górski, R. L., Lewandowska-Tomaszczyk, B., and Łaziński, M. (2008). Towards the National Corpus of Polish. In *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. In J. Levická and R. Garabík, editors, *Proceedings of Slovko 2009: Fifth International Conference on NLP, Corpus Linguistics, Corpus Based Grammar Research, 25–27 November 2009, Smolenice/Bratislava, Slovakia*, 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.