Which XML standards for multilevel corpus annotation?

Adam Przepiórkowski[1,2] and Piotr Bański[2]

[1] Institute of Computer Science PAS, ul. Ordona 21, 01-237 Warszawa, Poland
adamp@ipipan.waw.pl

[2] University of Warsaw, Krakowskie Przedmieście 26/28, 00-927 Warszawa, Poland
pkbanski@uw.edu.pl

Abstract. The paper attempts to answer the question: *Which XML standard(s) should be used for multilevel corpus annotation?* Various more or less specific standards and best practices are reviewed: TEI P5, XCES, work within ISO TC 37 / SC 4, TIGER-XML and PAULA. The conclusion of the paper is that the approach with the best claim to following text encoding standards consists in 1) using TEI-conformant schemata that are 2) designed in a way compatible with other standards and data models.

Keywords: corpus encoding, TEI, linguistic annotation, XML, ISO TC 37 / SC 4, TIGER-XML, 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.

One such project is the international project KYOTO (Knowledge Yielding Ontologies for Transition-based Organization; `http://www.kyoto-project.org/`), involving 11 institutions from Europe and Asia. Another is the much smaller National Corpus of Polish project (Pol. *Narodowy Korpus Języka Polskiego*; NKJP; `http://nkjp.pl/`; [22, 23]) involving 4 Polish institutions. What these two very different projects have in common is the strong emphasis on the conformance with current XML standards in LR encoding. It is interesting that this common objective gives rise to very different practices in these projects.

The aim of this paper is to present the way the National Corpus of Polish (henceforth, NKJP) attempts to follow standards and best practices in encoding multiple layers of linguistic annotation. The comparison of XML encoding schemata and underlying data models is a tedious and time-consuming task, and we hope that the following discussion will help other developers of LRs choose the standard best suited for their needs.

2 Requirements

NKJP is a project carried out in 2008–2010, aiming 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.

3 Standards and best practices

The three text-encoding standards and best practices listed in a recent CLARIN short guide ([6])[3] are: standards developed within ISO TC 37 / SC 4, the Text Encoding Initiative (TEI) guidelines and the XML version of the Corpus Encoding Standard (XCES). The following three subsections describe the current status of these standards, with two additional common practices briefly characterised in the subsequent subsections.

3.1 ISO TC 37 / SC 4

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] See also [3].

Acronyms of various standards potentially applicable in NKJP, as well as their current status (as given at `http://www.iso.org/`) and the latest publication freely available from `http://www.tc37sc4.org/`, are listed in Table 1.

standard	stage	version	available
FSR	IS	ISO/DIS 24610-1	2005-10-20
FSD	DIS	ISO/CD 24610-2	2007-05-03
WordSeg1	DIS	ISO/CD 24614-1	2008-06-24
MAF	DIS	ISO/CD 24611	2005-10-15
SynAF	DIS	ISO/CD 24615	2009-01-30
LAF	DIS	ISO/WD 2461[2]	2008-05-12

Table 1. Relevant ISO standards

The first two standards are concerned with feature structure representation (FSR) and declaration (FSD). WordSeg1 defines basic concepts and very general principles of word segmentation in diverse languages. The Morphosyntactic (MAF) and Syntactic (SynAF) Annotation Frameworks are specifications of the representation of wordform and syntactic (both constituency and dependency) information, respectively. Finally, the Linguistic Annotation Framework (LAF) defines a general abstract pivot format to which all levels of linguistic information may be mapped. Currently, only FSR is an actual published standard (ISO 24610-1).

3.2 TEI

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 primary data, with an active community, detailed guidelines ([4]) 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.

3.3 XCES

The Corpus Encoding Standard ([12, 10]), a corpus-centred offshoot of TEI, was developed within the Expert Advisory Group on Language Engineering Standards (EAGLES) project, and subsequently translated from SGML to XML ([11]). The resulting XCES specifies the encoding for primary data, for morphosyntactic annotation, and for alignment of parallel corpora. It also provides general feature structure mechanisms for the representation of other levels of information.

Although there are various resources and projects following XCES, including the IPI PAN Corpus of Polish (`http://korpus.pl/`; [19]), the standard apparently has not been modified since 2003; `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*. There are two sets of XML schemata, given as XML Schema (apparently last updated in 2003) and as DTD (apparently older), specifying different XML formats.

3.4 TIGER-XML

TIGER-XML ([16]) is a *de facto* standard for XML annotation of treebanks (syntactically annotated corpora). It is well documented and exemplified, it has been adopted in various projects, and it was the starting point for SynAF. In this schema, each sentence is represented as a `<graph>` consisting of `<terminals>` and `<nonterminals>`, where `<terminals>` is a list of `<t>`erminals (with orthographic, morphosyntactic and other information represented in attributes), and `<nonterminals>` is a list of `<nt>` syntactic nodes. Within each node, `<edge>`s link to immediate constituents (`<t>`s or `<nt>`s). Additional secondary edges (`<secedge>` elements within `<nt>`) may be used to represent co-reference information.

There is a treebank search engine working on TIGER-XML corpora, TIGERSearch ([15, 14]), 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 ([13, 20]) format.

3.5 PAULA

PAULA (Ger. *Potsdamer AUstauschformat für Linguistische Annotation*; [7]), 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.[4]

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:[5]

```
<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>
```

[4] See [8] for references to other such largely graph-based encodings.
[5] This is a modification of an example from [7].

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, `<feat>` elements associate markables with feature values.

4 Discussion

Of the *de facto* and purported standards described above, the first to be rejected is XCES, as 1) it has specific recommendations only for the linguistic level of morphosyntactic annotation, 2) the general feature structure mechanisms envisaged for other levels are different from FSR, an established ISO standard, 3) XCES includes no mechanisms for discontinuity, 4) or alternatives, and 5) there is a potential for confusion regarding the version of the standard. XCES was derived from TEI version P4, but it has not been updated to TEI P5 so far.

Apart from the TEI-derived XCES, TEI P5 is the only standard which includes detailed specifications for the encoding of metadata and text structure, so its deployment for these levels, as well as for text segmentation into sentences, is uncontroversial.

At the layers of word-level segmentation and morphosyntactic representation, the proposed ISO standards WordSeg1 and MAF are relevant. WordSeg1 provides general principles of word segmentation, and its main rule — that word segmentation should be lexicon-driven — is followed in NKJP.[6] MAF offers specific recommendations for the encoding — within a single XML file — of what we consider to be three layers: fine-grained segmentation, morphosyntactic analysis, and syntactic words. For this reason MAF cannot be applied verbatim in the project described here, and a more general stand-off representation must be adopted. The specific XML encodings proposed in §§5.2–5.4 may be easily mapped into MAF.

For the syntactic level, either the specific TIGER-XML encoding or the more general SynAF model may be employed. In fact, TIGER-XML is a concrete instantiation of SynAF. Unfortunately, TIGER-XML assumes that both terminal and non-terminal nodes are present within the same XML `<graph>` element, while in NKJP they should be separated, as there are two different and potentially conflicting syntax-like levels (syntactic groups and named entities) that

[6] The dictionary used in the project is a new version of Morfeusz ([26]), encoding the data of the *Słownik gramatyczny języka polskiego* ('Grammatical dictionary of Polish'; [24]). Occasionally, in well-defined cases, this general rule is in conflict with the principle of bound morpheme ("If a bound morpheme is attached to a word, then the result is a word.").

refer to the same word level. In §5.5 we propose a stand-off encoding inspired by (and mappable to) TIGER-XML, satisfying the general SynAF model.

None of the above standards provides specific mechanisms for representing word senses. In §5.6, we propose encoding analogous to that of morphosyntactic information, but implementing a mechanism of referring to particular entries within a sense dictionary, reminiscent of the `@entry` attribute in MAF.

Wherever there are no specific schemata for particular linguistic levels, general graph and feature structure representation mechanisms could be used as proposed, e.g., in LAF, and implemented, e.g., in PAULA. We follow this general approach, and the encodings proposed in the following sections are compatible with it. However, at this stage, the proposed ISO standards are still under development, with LAF and SynAF proposing only very general data models rather than specific solutions. Being aware of past efforts of developing annotation schemata which would "adhere as much as possible to the proposals for the new ISO/TC 37/SC 4 standard for linguistic resources" ([18]), but which do not adhere to them anymore as those proposed standards evolved, we decided to rely on established rather than proposed standards.

Two such general standards are TEI and PAULA; although the former is not generally thought of as a graph-encoding formalism, its reference mechanisms can be used to represent graphs in a way not less straightforward than that implemented in PAULA. In the end, we chose TEI P5 as the general encoding standard in NKJP also for a number of other reasons: 1) for primary data and metadata levels there is no real alternative to TEI, 2) TEI implements the ISO FSR standard, which can be used for the representation of linguistic content, as proposed in LAF (while PAULA introduces its own feature mechanism), 3) TEI is much more firmly established as a *de facto* standard for text encoding, with a much larger user base.

This approach is radically different from that adopted in the KYOTO project, mentioned in §1, where the approach of maximal adherence to established and proposed ISO ISO TC 37/SC 4 standards is assumed. This approach is justified to the extent that one of the main emphases of the project is the encoding of semantic dictionaries, and it relies in this regard on the established ISO 24613 standard (Lexical Markup Framework). Nevertheless, certain tensions resulting from the attempts to follow other, less developed ISO standards are visible in [1], where section 4, first describing MAF and SynAF on over 20 pages, ends with the following statement (p. 36):

We decided to remove MAF and SYNAF from the system design. Instead of that, we added to the KAF [Kyoto Annotation Framework] format some syntactic layers, thus representing among the different KAF levels also the morphological and syntactic levels. Basic motivation for that were that MAF is not finalized and complete, and that current documents are not consistent. Moreover SYNAF contains a lot of information that we do not need and, embedding representation of data into the original text documents, it complicates the representation and manipulation of information.

5 Standards in NKJP

For reasons discussed above, TEI P5 has been adopted as the main standard in NKJP. However, TEI is a rich toolbox, providing a variety of tools to address particular problems. Whenever there is a choice, an attempt has been made to select a solution isomorphic with other proposed, official and *de facto* standards.

5.1 Metadata, primary data and structure

The CLARIN short guide on metadata ([5]) makes the following recommendation: *We recommend using... (1) IMDI and its special profiles including TEI elements or (2) OLAC*, and later adds: *Also components and profiles will be offered that contain IMDI, TEI and OLAC specifications to take care of the already existing metadata records.* Hence, the use of TEI headers is in line with current best practices, and natural for LRs otherwise represented according to the TEI Guidelines. Apart from a TEI header for each text (`header.xml`), there is a general TEI corpus header, describing NKJP as a whole (`NKJP_header.xml`).

There is also no viable alternative to TEI for the representation of primary data and text structure. Texts are acquired for NKJP from a variety of sources, including previous participating corpora, publishers, Internet, media, original recordings of spontaneous conversations. They come with different kinds of structural information and different front and back matters. Some are divided into paragraphs or paragraph-like blocks, others into conversation turns. TEI Guidelines provide well-defined elements for all these situations.

TEI P5 encoding of metadata, primary data and structural information, as employed in the National Corpus of Polish, is presented in detail in [21]. The outline of `text_structure.xml`, containing a single text and any structural annotation, is as follows, with `<front>` and `<back>` (matter) elements optional:

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

In the case of written texts, the element `<body>` contains possibly nested `<div>` elements, expressing the overall structure of the text and containing `<p>`aragraphs (or paragraph-like anonymous blocks, `<ab>`). For spoken data, `<body>` consists of `<u>`tterances.

5.2 Segmentation

Within any `ann_segmentation.xml` file, the `<body>` element contains a sequence of `<p>`, `<ab>` or `<u>` elements mirroring those found in the `<body>` of the corresponding `text_structure.xml`. The parallelism is expressed via TEI `@corresp` attributes on these elements; their values refer to the corresponding elements in `text_structure.xml`. Any other structural markup is not carried over to this or other linguistic levels.

Each paragraph or utterance is further divided into `<s>`entences and even further into `<seg>`ments which define the span of each segment, by providing offsets to an appropriate element in `text_structure.xml`.[7] Each such `<seg>` element bears the implicit attribute `@type="token"`.

```
<seg xml:id="segm_1.1-seg"
 corresp="text_structure.xml#↩
 string-range(txt_1.1-p,0,6)"/>
```

5.3 Morphosyntax

The overall structure of `ann_morphosyntax.xml`, down to the level of `<seg>` (also implicitly marked as `@type="token"`), is identical to that of `ann_segmentation.xml`, with each `<seg>` referring — via the value of `@corresp` — to the corresponding segment at the segmentation level. Within `<seg>`, however, a feature structure — encoded in conformance with the FSR ISO standard — represents information about all morphosyntactic interpretations of a given segment, as well as about the tool used to disambiguate between them and the result of the disambiguation. For example, the logical structure of the content of a `<seg>` representing the noun *komputer* (singular, inanimate masculine, nominative or accusative) may be represented as follows:[8]

$$\begin{bmatrix} \textit{morph} & \\ \textsc{orth} & \text{komputer} \\ \textsc{interps} & \begin{bmatrix} \textit{lex} & \\ \textsc{base} & \text{komputer} \\ \textsc{ctag} & \text{subst} \\ \textsc{msd} & \text{sg:nom:m3} \vee \boxed{1}\text{sg:acc:m3} \end{bmatrix} \\ \textsc{disamb} & \begin{bmatrix} \textit{tool\_report} & \\ \textsc{tool} & \text{Anotatornia} \\ \textsc{date} & \text{2009-07-03 00:21:17} \\ \textsc{resp} & \text{PK + AA} \\ \textsc{choice} & \boxed{1} \end{bmatrix} \end{bmatrix}$$

Note that the names of features BASE, CTAG and MSD are taken from XCES. The value of INTERPS may actually be a list of feature structures, representing interpretations differing in base form (BASE) or grammatical class (CTAG). In

[7] Two complexities concerning alternative segmentations and information about boundedness of segments are discussed — and solutions are proposed — in [2].

[8] In this case manual disambiguation was performed by two annotators, anonymised here as PK and AA, with the help of a tool called Anotatornia.

cases where interpretations differ only in morphosyntactic description (MSD), they are listed locally, as alternative values of MSD. Hence, it is the value of MSD that is used for the disambiguation information within DISAMB|CHOICE.

5.4 Syntactic words

Word segmentation in the sense of the previous two levels, as produced by a morphological analyser used in NKJP, is very fine-grained: segments never contain spaces, and sometimes orthographic ("space-to-space") words are broken into smaller segments. For this reason an additional level is needed that will contain multi-token words, e.g., analytical tense forms of verbs. It is this level that corresponds most closely to MAF. However, while MAF assumes that `<token>`s and `<wordForm>`s reside in the same file (with `<token>` perhaps referring to primary data in a different file), we need a stand-off encoding referring to `ann_morphosyntax.xml`.

Down to the `<s>`entence level, `ann_words.xml` follows the same design as other levels, and links its `<s>` elements to those in `ann_morphosyntax.xml`, again via `@corresp`. Each sentence at this level is a list of `<seg>`ments of `@type="word"` covering the whole original sentence. In the default case, a `<seg>`ment at this level will be co-extensive with a `<seg>` at the lower level, but it may also correspond to a possibly discontinuous list of such token-level `<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.

One way to represent such syntactic words in TEI is given schematically below. The feature structure `<fs>` contains information about the lemma and the morphosyntactic interpretation of the word, similarly to the information at the morphosyntactic levels, but without ambiguities. Segments in `ann_morphosyntax.xml` (and possibly syntactic words in `ann_words.xml`) within the given word are referenced via the `<ptr>` element.

```
<seg xml:id="word13">
 <fs>...</fs>
 <ptr target="ann_morphosyntax.xml#seg15"/>
 <ptr target="ann_morphosyntax.xml#seg16"/>
 <ptr target="ann_morphosyntax.xml#seg18"/>
</seg>
```

5.5 Named entities and syntactic groups

Files representing the following two levels, `ann_named.xml` for named entities (NEs) and `ann_groups.xml` for syntactic groups, also have the same overall structure down to the `<s>` level, but within each sentence only the information pertinent to the current level is represented, so, in particular, some `<s>` elements within `ann_named.xml` may be empty, if the relevant sentences do not contain any named entities. Both levels refer independently to the level of syntactic words.

Within ann_groups.xml, each sentence is a sequence of <seg>ments of @type="group" structured in a way analogous to the word-level <seg> elements described above: they consist of a feature structure describing the syntactic group, as in the following simplified example. Note that the @type attribute of <ptr> defines the kind of relation between the node and its immediate constituent; note also that <ptr> elements have @xml:id values and, hence, may be referenced from within the <fs> description of the group.

```
<seg xml:id="group4">
 <fs>...</fs>
 <ptr xml:id="id1" type="head"
  target="ann_words.xml#word10"/>
 <ptr xml:id="id2" type="nonhead"
  target="ann_words.xml#word12"/>
 <ptr xml:id="id3" type="nonhead"
  target="#group3"/>
</seg>
```

The representation of NEs is analogous, with the following differences: 1) the implicit value of @type is "named" instead of "group", 2) different information is represented within the <fs> description; this includes the type of the named entity, as well as the base form of the NE, which, obviously, does not need to be a simple concatenation of base forms of words within the NE, 3) there seems to be no need for the @type attribute within <ptr>.

5.6 Word senses

Within NKJP, a limited number of semantically ambiguous lexemes will be disambiguated.[9] In a manner analogous to the morphosyntactic level, each <s> contains a sequence of token-level <seg>ments, with @corresp references to <seg>ments in ann_segmentation.xml.[10] Each <seg> contains a feature structure with a reference to the appropriate sense in an external word sense inventory, e.g.:

```
<seg xml:id="seg2"
 corresp="ann_segmentation.xml#seg17">
 <fs type="sense">
  <f name="sense" fVal="NKJP_WSI.xml#sam.2"/>
 </fs>
</seg>
```

In a way analogous to the two levels described in the preceding subsection, only those segments are represented here which were semantically disambiguated, so some <s> elements will be empty.

[9] See also [17] in these proceedings.

[10] This is a technical decision; in the future, the word sense level may be changed to reference syntactic words rather than segments.

6 Conclusion

For each specific TEI P5 solution presented above there are other ways of representing the same information in a way conformant with the TEI P5 Guidelines. For example, instead of recycling the `<seg>` element with different `@type` values, TEI elements such as `<w>` (for words), `<phr>` and `<cl>` (for syntactic groups), and even `<persName>`, `<geogName>`, `<orgName>` and `<date>` (for various kinds of named entities) could be used at different levels. Instead of using `<ptr>` links, nested structures could be represented straightforwardly via the nesting of XML elements, or — much less straightforwardly — as feature structures ([25]), etc.

The encoding proposed in this paper was designed with the view of maximising compatibility with other standards, whether sanctioned by ISO or *de facto* in use. It is directly mappable to specific encodings such as TIGER-XML and PAULA, and it is an instantiation of sometimes rather abstract models developed within ISO TC 37 / SC 4.

We conjecture that — given the stability, specificity and extensibility of TEI P5 and the relative instability and generality of some of the other proposed standards — this approach is currently the optimal way of following corpus encoding standards.

References

1. Aliprandi, C., Neri, F., Marchetti, A., Ronzano, F., Tesconi, M., Soria, C., Monachini, M., Vossen, P., Bosma, W., Agirre, E., Artola, X., de Ilarraza, A.D., Rigau, G., Soroa, A.: Database models and data formats (2009), KYOTO Deliverable NR. 1/WP NR. 2, Version 3.1, 2009-01-31
2. Bański, P., Przepiórkowski, A.: 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. pp. 64–67. Singapore (2009)
3. 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., Wittenburg, P.: Standardisation action plan for Clarin (2009), state: Proposal to CLARIN Community; August 2009
4. Burnard, L., Bauman, S. (eds.): TEI P5: Guidelines for Electronic Text Encoding and Interchange. Oxford (2008), `http://www.tei-c.org/Guidelines/P5/`
5. Component metadata: A CLARIN shortguide (2009), `http://www.clarin.eu/documents`
6. Standards for text encoding: A CLARIN shortguide (2009), `http://www.clarin.eu/documents`
7. Dipper, S.: Stand-off representation and exploitation of multi-level linguistic annotation. In: Proceedings of Berliner XML Tage 2005 (BXML 2005). pp. 39–50. Berlin (2005)
8. Dipper, S., Hinrichs, E., Schmidt, T., Wagner, A., Witt, A.: Sustainability of linguistic resources. In: Hinrichs, E., Ide, N., Palmer, M., Pustejovsky, J. (eds.) Proceedings of the LREC 2006 Workshop on Merging and Layering Linguistic Information. pp. 14–18. ELRA, Genoa (2006)
9. ELRA: Proceedings of the Second International Conference on Language Resources and Evaluation, LREC 2000 (2000)

10. Ide, N.: Corpus Encoding Standard: SGML guidelines for encoding linguistic corpora. In: Proceedings of the First International Conference on Language Resources and Evaluation, LREC 1998. pp. 463–470. ELRA, Granada (1998)
11. Ide, N., Bonhomme, P., Romary, L.: XCES: An XML-based standard for linguistic corpora. In: LREC [9], pp. 825–830
12. Ide, N., Priest-Dorman, G.: Corpus encoding standard (1995), `http://www.cs.vassar.edu/CES/`, accessed on 2009-08-22
13. Janus, D., Przepiórkowski, A.: Poliqarp: An open source corpus indexer and search engine with syntactic extensions. In: Proceedings of the ACL 2007 Demo and Poster Sessions. pp. 85–88. Prague (2007)
14. König, E., Lezius, W., Voormann, H.: TIGERSearch 2.1: User's Manual. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart (2003)
15. Lezius, W.: TIGERSearch — ein Suchwerkzeug für Baumbanken. In: Busemann, S. (ed.) Proceedings der 6. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2002). Saarbrücken (2002)
16. Mengel, A., Lezius, W.: An XML-based encoding format for syntactically annotated corpora. In: LREC [9], pp. 121–126
17. Młodzki, R., Przepiórkowski, A.: The WSD development environment. In: Vetulani, Z. (ed.) Proceedings of the 4th Language & Technology Conference. pp. 185–189. Poznań, Poland (2009)
18. Pianta, E., Bentivogli, L.: Annotating discontinuous structures in XML: the multiword case. In: Proceedings of the LREC 2004 Workshop on XML-based Richly Annotated Corpora. pp. 30–37. ELRA, Lisbon (2004)
19. Przepiórkowski, A.: The IPI PAN Corpus: Preliminary version. Institute of Computer Science, Polish Academy of Sciences, Warsaw (2004)
20. Przepiórkowski, A.: Powierzchniowe przetwarzanie języka polskiego. Akademicka Oficyna Wydawnicza EXIT, Warsaw (2008)
21. Przepiórkowski, A., Bański, P.: XML text interchange format in the National Corpus of Polish. In: Goźdź-Roszkowski, S. (ed.) The proceedings of Practical Applications in Language and Computers PALC 2009. Peter Lang, Frankfurt am Main (2009)
22. Przepiórkowski, A., Górski, R.L., Lewandowska-Tomaszczyk, B., Łaziński, M.: Towards the National Corpus of Polish. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC 2008. ELRA, Marrakech (2008)
23. Przepiórkowski, A., Górski, R.L., Łaziński, M., Pęzik, P.: Recent developments in the National Corpus of Polish. In: Levická, J., Garabík, R. (eds.) NLP, Corpus Linguistics, Corpus Based Grammar Research: Proceedings of the Fifth International Conference, Smolenice, Slovakia, 25–27 November 2009. pp. 302–309. Tribun, Brno (2009)
24. Saloni, Z., Gruszczyński, W., Woliński, M., Wołosz, R.: Słownik gramatyczny języka polskiego. Wiedza Powszechna, Warsaw (2007)
25. Witt, A., Rehm, G., Hinrichs, E., Lehmberg, T., Stemann, J.: SusTEInability of linguistic resources through feature structures. Literary and Linguistic Computing 24(3), 363–372 (2009)
26. Woliński, M.: Morfeusz — a practical tool for the morphological analysis of Polish. In: Kłopotek, M.A., Wierzchoń, S.T., Trojanowski, K. (eds.) Intelligent Information Processing and Web Mining, pp. 511–520. Advances in Soft Computing, Springer-Verlag, Berlin (2006)