XML Text Interchange Format in the National Corpus of Polish[*]

Adam Przepiórkowski[*,†] and **Piotr Bański**[†]

[*]Institute of Computer Science, Polish Academy of Sciences
and [†]University of Warsaw

Abstract

The aim of this paper is to describe and justify the XML encoding of texts within the National Corpus of Polish. Basic text encoding, rather than linguistic annotation, is considered here: the encoding of the primary data, the structural markup and the metadata. A set of schemata conformant with the Text Encoding Initiative Guidelines P5 is presented.

Keywords: Text Encoding Initiative, TEI P5, National Corpus of Polish, NKJP, metadata, TEI header, structural markup, primary data, XML, Polish.

1. Introduction

National Corpus of Polish (Pol. *Narodowy Korpus Języka Polskiego*; NKJP; `http://nkjp.pl/`) is a project carried out in 2008–2010, involving 4 Polish institutions: Institute of Computer Science of the Polish Academy of Sciences (coordinator), Institute of Polish Language of the Polish Academy of Sciences, University of Łódź and Polish Scientific Publishers PWN.[1] Each of these institutitions contributes texts from their own corpora, and each — apart from the coordinator — acquires new texts for the National Corpus of Polish (NKJP, henceforth): books, newspapers and magazines, blogs, transcripts of spoken data, etc. All these texts are imported into two very different search engines available in NKJP (cf. the "Demo" link at `http://nkjp.pl/`).

Obviously, before NKJP texts can be indexed or automatically processed by any other tools they must be converted to a common interchange format. Such interchange

[*]Research funded in 2007–2010 by a research and development grant from the Polish Ministry of Science and Higher Education.

[1]A programmatic description of the project may be found in Przepiórkowski et al. 2008, and more recenent developments are presented in Przepiórkowski et al. 2009.

format should allow for the representation of various types of texts mentioned above, and also for the encoding of various kinds of metadata and structural information. The only text encoding standard sufficiently versatile to meet these requirements is TEI P5, presented in the Guidelines of the Text Encoding Initiative (TEI; Burnard & Bauman 2008; `http://www.tei-c.org/`). It is not an official ISO standard, but a mature and very specific XML-based *de facto* standard for text encoding in the humanities, with a rich user base and supporting tools.

The reason for continuing this paper beyond the previous paragraph is that TEI is a large treasure trove of solutions, rather than a lean and highly focussed formalism, and a particular text encoding schema must still be designed by choosing the most appropriate mechanisms from the TEI toolbox and — in rare specific cases — by introducing new XML elements or attributes. The aim of this paper is to present and document one such particular schema, developed within NKJP. As there are few well-documented TEI corpora around, and hardly any corpora following the current P5 version of the TEI Guidelines (substantially differing from the previous TEI versions), we hope that this presentation will facilitate the development of other TEI P5 corpora.

The remainder of the paper starts, in §2, with a presentation of the NKJP corpus header, i.e., an XML document containing metadata pertaining to the National Corpus of Polish as a whole. The representation of text headers, i.e., metadata for particular texts, is described in §3. The ensuing section, §4, makes clear the overall structure of a corpus text and the place of both kinds of metadata in that structure. This section also sketches the representation of structural and typographical distinctions within texts. Although, within NKJP, texts are also annotated at various linguistic levels, this paper does not deal with such linguistic annotation — see Przepiórkowski & Bański 2009 for an overview and Bański & Przepiórkowski 2009 for a discussion of some technical issues. Finally, §5 concludes the paper.

2. Corpus Header

Following the TEI Guidelines, the NKJP corpus header consists of 4 sections contained in the `<teiHeader xml:lang="en" type="corpus">` element: `<fileDesc>`,

<profileDesc>, <encodingDesc> and <revisionDesc>.

Two of these have very simple structure. First, <profileDesc> identifies the main languages used in the TEI encoding of texts and metadata, and it is cited in its entirety below:

```
<profileDesc>
 <langUsage>
  <language ident="pl">Polish</language>
  <language ident="en">English</language>
 </langUsage>
</profileDesc>
```

The values of @ident attributes may be used for any element to specify the language of the content of that element. In fact, the xml:lang="en" specification in the <teiHeader> element is inherited by other elements in the header, unless explicitly overridden by xml:lang="pl", thus making English the default language of the NKJP header.

Another simple and homogeneous section is <revisionDesc>: it contains a sequence of <change> statements like the following:

```
<change who="#adamp" when="2009-08-01">Added <gi>profileDesc</gi>.</change>
```

The <fileDesc> section contains 4 subsections. The first, <titleStmt>, specifies the name of the corpus and describes the responsibility of various institutions and persons involved in its creation. One such responsibility statement is referenced by who="#adamp" in the example above, another may look as follows:

```
<respStmt>
 <persName xml:id="bansp">Piotr Bański</persName>
 <resp>initial design of various XML schemata</resp>
</respStmt>
```

The other three subsections of <fileDesc> are: <editionStmt> — a brief statement concerning the stability of the current version of NKJP, <publicationStmt> — defining availability and distribution of NKJP, and <sourceDesc> — specifying the origin of texts in general terms (specific source descriptions are contained in the headers of particular texts).

Finally, <encodingDesc> characterizes NKJP in various ways, e.g., <projectDesc> repeats the description of the project given at http://nkjp.pl/, <samplingDecl> says that

Whole texts are included, whenever possible and provides some information on text structure, as discussed in § 4, and `<editorialDecl>` briefly discusses anonymisation of spoken data and other editorial interventions in NKJP texts.

While these subsections contain free-text statements, many other `<encodingDesc>` subsections are more structured. Perhaps the most important are `<classDecl>` subsections, which specify text classifications referenced in particular text headers. For example, one of the ways in which NKJP texts are classified is according to the Universal Decimal Classification, so the following declaration is present in the corpus header:

```
<classDecl>
 <taxonomy xml:id="ukd">
  <bibl>
   <title xml:lang="pl">Uniwersalna Klasyfikacja Dziesiętna</title>
   <title xml:lang="en">Universal Decimal Classification</title>
   <edition>UDC-P058</edition>
  </bibl>
 </taxonomy>
</classDecl>
```

Within a text header (cf. § 3 below), a reference to this classification may be made as follows: `<classCode scheme="#ukd">821.162.1-3</classCode>`. Similarly, in order to control the good balance of the corpus with respect to genres, a taxonomy of text types is defined; its fragment is presented below:

```
<classDecl>
 <taxonomy xml:id="taxonomy-NKJP-type">
 <!-- ... --->
   <category xml:id="typ_lit_proza">
    <desc xml:lang="pl">proza</desc>
    <desc xml:lang="en">prose</desc>
   </category>
   <category xml:id="typ_lit_poezja">
    <desc xml:lang="pl">poezja</desc>
    <desc xml:lang="en">poetry</desc>
   </category>
   <category xml:id="typ_lit_dramat">
    <desc xml:lang="pl">dramat</desc>
    <desc xml:lang="en">drama</desc>
   </category>
 <!-- ... --->
 </taxonomy>
</classDecl>
```

Again, the type of a particular text may be defined by referencing one of the categories defined in such a classification.

The final <encodingDesc> subsection in the NKJP header to be mentioned here[2] is <nkjp:fsLib>. As the namespace prefix nkjp suggests, this element is not defined by TEI but introduced within NKJP for a specific purpose.

TEI specifications contain the ISO standard on the XML representation of feature structures (ISO:24610-1 2005) and, within NKJP, feature structures are used for representing various types of linguistic annotation (Przepiórkowski & Bański 2009). The standard makes it possible to define feature structure libraries containing, e.g., feature structures representing complex morphosyntactic information. Such feature structures may subsequently be referenced from an appropriate linguistic layer by their identifier, thus simultaneously increasing readability and compactness of linguistic representations. Curiously, according to TEI specifications it is not possible to define such feature structure libraries in a header, which seems to be the most natural place for such libraries: once they are in the corpus header, they may be referenced in a way analogous to how particular categories defined within <classDecl> are referenced for classification. Hence the need for the project-specific element <nkjp:fsLib>.

The presence of such nkjp:... elements and attributes is what makes the NKJP schema TEI conformant in a weaker sense: it is a TEI Extension rather than TEI Conformant (with a capital 'C'), as defined in Burnard & Bauman 2008, § 23.3. Nevertheless, there are only a few conservative and well-justified modifications of this kind in the NKJP schema presented here, so it may be regarded as a "nearly" TEI Conformant TEI Extension.

3. Text Header

Each NKJP text is represented as a number of XML files, two of which are relevant here: header.xml and text_structure.xml.[3] The structure of the latter will be presented in § 4. The structure of text headers, header.xml, is similar to that of the corpus header:

[2] Two other such subsections present in the NKJP header are <tagsDecl> and <refsDecl>; see appropriate links in http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-encodingDesc.html for a description of their role in <encodingDesc>.

[3] Other files contain various layers of linguistic annotation, as described in Przepiórkowski & Bański 2009.

the <teiHeader> element, implicitly marked as type="text" here, contains three sections: <fileDesc>, <profileDesc> and <revisionDesc>. The last section, <revisionDesc>, is fully analogous to that of the corpus header and contains a sequence of <change> elements describing modifications to any of the files representing the text and its annotation.

On the other hand, <profileDesc> differs from that of the corpus header and it comprises one element, <textClass>, which contains classifiers of the text, referencing appropriate taxonomies posited in the corpus header. For example, the content of <profileDesc> for Manuela Gretkowska's novel *Namiętnik* may look as follows:

```
<profileDesc>
 <textClass>
  <classCode scheme="#ukd">821.162.1-3</classCode>
  <keywords scheme="#bn">
   <list>
    <item>Opowiadanie polskie -- 20 w.</item>
   </list>
  </keywords>
  <catRef scheme="#taxonomy-NKJP-type" target="#typ_lit_proza"/>
  <catRef scheme="#taxonomy-NKJP-channel" target="#kanal_ksiazka"/>
 </textClass>
</profileDesc>
```

References to 4 classification schemes are made here: two external to the project (the Universal Decimal Classification mentioned above and the classification of the Polish National Library, cf. #bn), and two internal (text type: #typ_lit_proza, i.e., prose, and text channel: #kanal_ksiazka, i.e., book).

Finally, <fileDesc> contains a variety of information about the text: its title in NKJP (e.g., "TEI P5 encoded version of "Namiętnik""), bibliographic information about the source of the text (title, author, publisher, etc.), a note about the origin of the text in NKJP (e.g., "<note type="text_origin">IPI PAN Corpus</note>"), the original header, if available, of the text as it was defined in the corpus from which the text is inherited, as well as a <publicationStmt>, exemplified below:

```
<publicationStmt nkjp:subcorpus="balanced">
 <availability status="restricted">
  <p>For all NKJP purposes.</p>
 </availability>
</publicationStmt>
```

The `@nkjp:subcorpus` attribute shown above is another example of a non-Conformant modification of TEI, needed here in order to represent the information about the target NKJP subcorpus for which the text was acquired.

For transcripts of spoken data, the text header may also contain information about the person responsible for transcription (encoded as `<respStmt>` within `<fileDesc>`), various kinds of information about the source of the text (different from written texts, because here the source is a recording rather than a publication), as well as another element from the `nkjp` namespace, `<nkjp:topic>`, describing the topic of the conversation. Moreover, apart from `<textClass>`, `<profileDesc>` also contains a `<langUsage>` element specifying the level of formality of the conversation, `<particDesc>`, containing background information about participants in the conversation, as well as `<settingDesc>`, mentioning when and where the conversation took place; some of these elements specific to transcripts of spoken data are exemplified below:

```
<langUsage>
 <language ident="pl-x-formal"/>
</langUsage>

<nkjp:topic xml:lang="pl">Rozmowa o immunitecie Zbigniewa Ziobro,
 sytuacji w Gruzji i reakcji unii europejskiej na nią.</nkjp:topic>

<particDesc>
 <!-- ... --->
 <person xml:id="sp2" role="speaker">
  <persName>Zbigniew Ziobro</persName>
  <sex value="1">male</sex>
  <education xml:lang="pl">wyższe</education>
  <age>40</age>
  <residence>unknown</residence>
 </person>
 <!-- ... --->
</particDesc>

<settingDesc>
 <setting>
  <name type="place">TVP Info</name>
  <name type="voivodship" xml:lang="pl">mazowieckie</name>
  <date type="recorded" when="2008-09-02"/>
 </setting>
</settingDesc>
```

4. Text Structure

For any corpus document, `text_structure.xml` contains the actual text, as well as the structural markup of the document.

It is often considered best practice to have a read-only pure text file referenced by a stand-off file containing structural information. The main justification for this requirement is the need for an immutable text level. While stand-off annotation is used for all other NKJP layers, it has turned out to be impractical to separate primary data and structure. The reason for that is that corpus data are virtually never acquired as pure text, but almost always come with some markup already present: XML, HTML, or even implicit markup in Microsoft Word and OpenOffice files. Separating this markup from text for the sake of stand-off annotation (rather then converting it *in situ* into the appropriate TEI markup), and then logically combining them again for processing at later stages, would only generate unnecessary work. Hence, for the purposes of NKJP, it is `text_structure.xml` that is considered immutable.

The outline of `text_structure.xml`, containing a single text and any structural annotation, is as follows, with the `<front>` and `<back>` matter elements optional (always absent in transcripts of spoken data):

```
<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><!-- main text body --></body>
   <back><!-- back matter --></back>
  </text>
 </TEI>
</teiCorpus>
```

It should be noted that each text is a `<teiCorpus>` and logically includes not only the text header (`<xi:include href="header.xml"/>`), but also the entire corpus header (`<xi:include href="NKJP_header.xml"/>`).

For `<front>` and `<back>`, any structural elements defined in TEI P5 are allowed. Typically, within front matter there will be a title statement, possibly distinguishing between

the main title and the subtitle, as in the following example:

```
<docTitle>
 <titlePart type="main">Pieśni nędzy i zagłady</titlePart>
 <titlePart type="sub">Twórczość Mordechaja Gebirtiga w Salonie Poezji</titlePart>
</docTitle>
```

On the other hand, in NKJP, the content of <body> is constrained with respect to the range of possibilities offered by TEI P5. For spoken texts, only a sequence of <u>tterances (and perhaps <incident>s between them) may occur within <body>, as in the following fragment:

```
<body xml:id="txt_body">
 <u who="#sp3" xml:id="u1">ale zostaw to w ogóle dajcie buziaka
  przepraszam was laski</u>
 <u trans="overlap" who="#sp1" xml:id="u2">no dałyśmy sobie buziaka
  no</u>
 <!-- ... -->
</body>
```

For written texts, the main blocks are <p> (paragraph), <ab> (anonymous block, i.e., a paragraph-sized chunk of text exceptionally used for texts without division into paragraphs) and <head> (for headings starting a textual division at any level, e.g., for chapter and section titles). These blocks can be grouped into chapters, sections, subsections, etc., using the — possibly nested — elements, e.g.:

```
<body>
 <!-- ... -->
 <div type="chapter" n="1">
  <head>Rozdział 1 Skąd się biorą paradygmaty?</head>
  <p><!-- ... --></p>
  <p><!-- ... --></p>
  <p><!-- ... --></p>
 </div>
 <div type="chapter" n="2">
  <head>Rozdział 2 Świat według Pszczółki Mai</head>
  <p><!-- ... --></p>
  <p><!-- ... --></p>
  <p><!-- ... --></p>
 </div>
 <!-- ... -->
</body>
```

The content of the elements <u>, <p>, <ab> and <head> is (almost) pure text; the only XML elements which may appear there are: <gap> (to mark places where tables, pictures, etc., were removed from text), <hi>, with the obligatory attribute @rend specifying how the highlighted text is rendered (in written texts only), <lb> (always empty, to mark line breaks in poetry or in a motto; only in written texts) and two elements used in spoken texts for marking non-verbal events, <vocal> and <incident>.

These restrictions on <u>, etc., are caused by the fact that the content of these elements will be further marked linguistically and indexed by search engines, so any additional markup which would complicate the operation of these tools must be well justified. More specifically, the following XPath expressions define the elements to be further processed:

- `//body//(p|ab|u|incident|head)`
- `//front//titlePart`

In other words, the content of any <p>, <ab>, <u>, <incident> and <head> elements anywhere within <body>, as well as the content of <titlePart> anywhere within <front> will be marked linguistically and indexed by corpus search engines.

5. Conclusion

Given that TEI P5 Guidelines were released in November 2007, there are still very few publicly available or documented TEI P5 corpora around. One that provided inspiration for some of the solutions presented above is the corpus of the project "Communication in Slovene" (http://nl.ijs.si/ssj/), although currently it does not represent nearly as many different kinds of linguistic and structural information as NKJP.[4]

Because of this lack of a large base of well-documented TEI P5 corpora, and because of the daunting number of choices that TEI makes available in its 1350-page guidelines, the design of the TEI-based NKJP interchange format turned out to be a tedious and time-consuming task. We hope that this paper, together with the accompanying articles cited above, will facilitate the design of similar TEI P5 schemata in other corpus projects.

[4] Also, for a minor technical reason (adding new elements to the TEI namespace) it is not TEI conformant in any sense of TEI conformance.

Bibliography

Bański, P. & A. Przepiórkowski. (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*. Singapore, 64–67.

Burnard, L. & S. Bauman (Eds.) (2008) *TEI P5: Guidelines for electronic text encoding and interchange*. Oxford. `http://www.tei-c.org/Guidelines/P5/`

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

Przepiórkowski, A. & P. Bański. (2009) 'Which XML standards for multilevel corpus annotation?'. Unpublished manuscript, submitted to the 4th Language and Technology Conference, Poznań, 2009.

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

Przepiórkowski, A., Górski, R. L., Łaziński, M. & P. Pęzik. (2009) 'Recent developments in the National Corpus of Polish'. In: Levická, J. & R. Garabík (Eds.), *Proceedings of Slovko 2009: Fifth international conference on NLP, corpus linguistics, corpus based grammar research, 25–27 November 2009, Smolenice/Bratislava, Slovakia*. Brno: Tribun.