Stand-off TEI Annotation: the Case of the National Corpus of Polish

Piotr Bański
Institute of English Studies
University of Warsaw
Nowy Świat 4, 00-497 Warszawa, Poland
pkbanski@uw.edu.pl

Adam Przepiórkowski
Institute of Computer Science
Polish Academy of Sciences
Ordona 21, 01-237 Warszawa, Poland
adamp@ipipan.waw.pl

Abstract

We present the annotation architecture of the National Corpus of Polish and discuss problems identified in the TEI stand-off annotation system, which, in its current version, is still very much unfinished and untested, due to both technical reasons (lack of tools implementing the TEI-defined XPointer schemes) and certain problems concerning data representation. We concentrate on two features that a stand-off system should possess and that are conspicuously missing in the current TEI Guidelines.

1 Introduction

The present paper presents the National Corpus of Polish (NCP).[1] The project is a joint undertaking of a consortium consisting of institutions that created their own large corpora of Polish in the past (see (Przepiórkowski et al., 2008) for details); these corpora formed the initial data bank of the corpus. The intended size of the corpus is one billion (10^9) tokens and as such, at the time of completion in 2010, the NCP is going to be one of the largest corpora available, possibly the largest corpus featuring multiple levels of linguistic annotation of various kinds. Currently, a hand-verified one-million-token subcorpus is being completed, and a basic, automatically created 430-million-token demo is available online at http://nkjp.pl/.

The project uses an extended morphosyntactic tagset with several years of practical use behind it in one of the source corpora (cf. http://korpus.pl/) and an open-source query engine with a powerful, regex-based language and a graphical front-end.

Section 2 of this paper talks about the encoding format adopted for the corpus, section 3 presents its general architecture, and section 4 discusses the reasons for, and our implementation of, the suggested NCP enhancements to the TEI Guidelines.

[1] The Polish name of the corpus is *Narodowy Korpus Języka Polskiego*, hence the abbreviation NKJP, used in web addresses and namespace identifiers.

2 The encoding format: stand-off TEI

The Text Encoding Initiative (TEI Consortium, 2007) has been at the forefront of text annotation and resource interchange for many years. It has influenced corpus linguistic practices in at least three related ways. Firstly, the formalism itself, in the mature form, has been used to mark up linguistic corpora, e.g. the British National Corpus. An early application of the TEI, the Corpus Encoding Standard (CES; see http://www.cs.vassar.edu/CES/), together with its XML version, XCES (http://www.xces.org/), have served as *de facto* standards for corpus encoding in numerous projects. Finally, the experience gained in creating and using XCES (together with e.g. the feature-structure markup of the TEI) has served as a foundation for the Linguistic Annotation Format (LAF, Ide and Romary, 2007), within ISO TC37 SC4. LAF promises to provide a standard interchange format for linguistic resources of many diverse kinds and origins.

The relationship between the TEI (especially in its stand-off version) and the LAF is straightforward. Both are implemented in XML, which makes transduction between a rigorous TEI format and the LAF "dump" (pivot) format mostly a matter of fleshing out some data structures.

3 NCP – general architecture

Stand-off annotation is by now a well-grounded data representation technique, pioneered by the CES and continuing to be the foundation of the LAF. In short, it assumes that the source text in the corpus, ideally kept in an unannotated form and in read-only files, is the root of a possibly multi-file system of data descriptions (each description focusing on a distinct aspect of the

Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP 2009, pages 64–67,
Suntec, Singapore, 6-7 August 2009.

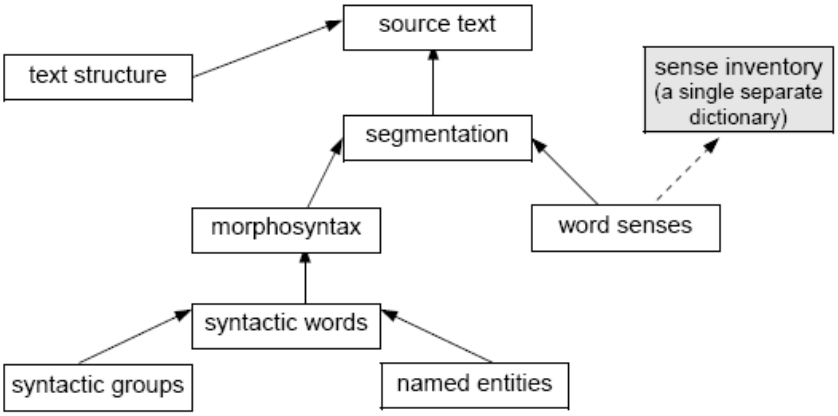

Figure 1: The logical data structure of the NCP

source data). The source text is typically accompanied by a level of primary segmentation, which may be the lowest-level XML layer of annotation. The other files form a possibly multi-leaved and multi-leveled hierarchy referencing either the level of primary segmentation, or higher order levels of description. The NCP follows these guidelines to the extent allowed by the TEI schema.

Each corpus text is kept in a separate directory together with the annotation files that reference it directly or indirectly, and with the header that is included by all these files. Contents of an example directory are shown below.

(1)
```
text.xml
header.xml
ann_morphosyntax.xml
ann_segmentation.xml
ann_structure.xml
```

All of these files contain TEI documents (or, in the case of header.xml, proper subsets thereof). They form a hierarchy of annotation levels, as presented in Figure 1. The text.xml file is the root, referenced by the layer of text structure (providing markup from the paragraph level upwards) and the layer of segmentation. The segmentation layer is further referenced by the layer of morphosyntactic information and word-sense annotation. The morphosyntactic level, in turn, is the basis for the level identifying syntactic words, which constitutes the foundation upon which the levels identifying syntactic chunks and named entities are built.

In text.xml, the normalized source text is divided in paragraph-sized chunks (enclosed in anonymous blocks, <ab>, to be further refined in the text-structure level of annotation).[2] It also includes two headers: the main corpus header, which encodes information relevant to all parts of the corpus, and the local header, which records the information on the particular text and its annotations.

The segmentation file provides what the LAF calls the base segmentation level that is further used as the basis for other kinds of annotation. It is implemented as a TEI document with <seg> elements that contain XInclude instructions (see example (4) in the next section). As such, it may serve both as a separate annotation layer or as a merged structure, after the inclusion directives are resolved. Crucially, in the latter case, which is the default with many parsers, the XPointer indexing information is lost. We shall come back to this issue in section 4.1.

The text-structure layer is defined similarly to the segmentation layer. Other annotation layers replace the mechanism of XInclude with XLink, in the way advocated by the XCES.

The morphosyntactic layer of annotation consists of a series of <seg> elements that contain TEI feature structures (i) providing basic information on the segment, (ii) specifying the possible interpretations as identified by the morphological analyser, and (iii) pointing at the morpho-

[2] Ideally, as mentioned above, the primary text should be stored without markup, and the segmentation layer should constitute the lowest-level XML document. This is exactly what the LAF-encoded American National Corpus does, requiring dedicated tools for merging plain text corpus files with the segmentation documents. Unfortunately, this is where we reach the technological boundary of the XInclude system: it is unable to reference substrings in a plain text file, due to a weakly motivated ban on the concurrent presence of @parse="text" attribute and the @xpointer attribute. We therefore enclose the source text in anonymous blocks (<ab>) that we can easily address with XPointers. An anonymous reviewer agrees that the lack of a single, immutable text file is a serious weakness of this system and notes that being able to derive plain text from markup is no remedy. This may constitute either a case for XLink, or an argument for lifting the @parse/@pointer ban.

syntactic description selected by the disambiguating agent.

The higher-order annotation layers also contain feature structures, which usually point at the selected segments of annotation layers that are one level lower, and identify their function within the given data structure.

4 Enhancements to the TEI stand-off recommendations

In this section, we first illustrate a case where the stand-off annotation system as advocated by the TEI loses information on the boundedness of segments, and then move on to illustrate a different issue stemming from the lack of a neutral bracket-like element in the TEI markup.

4.1 Identification of bound segments

Segmentation of Polish texts is not a trivial matter, partially because of the person-number enclitics – elements that can attach to almost any part of the clause, while being functionally related to the main verb. Segmenting them together with their hosts, apart from being a methodologically bad move, would greatly increase the complexity of the linguistic analysis built on top of such segmentations. The diamond in (2) below marks alternative positions where the 2nd Person Plural clitic (separated by a vertical bar) may appear. All of the resulting sentences have the same interpretation.

(2) Czemu|ście znowu♦ wczoraj♦ Piotra♦ gonili♦?
why|2pl again yesterday Piotr chased.prt
"Why did you chase Piotr yesterday again?"

Yet another group of segmentation problems concerns compounds, right-headed (3a) or coordinative (3b).

(3) a. żółto|czerwony materiał
yellow|red fabric
"yellowish red fabric"
b. żółto-czerwony materiał
"yellow and red fabric"

Inline markup of the above examples preserves information on which segment is bound (attached to the preceding one) or free-standing. This is due to the whitespace intervening between the <seg> elements in this kind of markup.

When, however, stand-off markup using the XInclude mechanism is applied here, complications arise. The segmental level of annotation with unresolved inclusions provides clear hints about the status of segments. This is due to XPointer offsets, as can be seen in (4) below, which is an example assuming that the adjective *żółto-czerwony* is the first word in an <ab> element bearing the @xml:id attribute set to "t1".[3]

(4)
```
<seg xml:id="segm_1.1-seg">
 <xi:include href="text.xml"
 xpointer="string-range(t1,0,5)"/></seg>
<seg xml:id="segm_1.2-seg">
 <xi:include href="text.xml"
 xpointer="string-range(t1,5,1)"/></seg>
<seg xml:id="segm_1.3-seg">
 <xi:include href="text.xml"
 xpointer="string-range(t1,6,8)"/></seg>
```

However, after inclusions are resolved, all of the offset information is lost, because all the @xpointer attributes (indeed, all the <xi:include> elements) are gone and all that remains is a sequence of <seg> elements such as `<seg>żółto</seg><seg>-</seg><seg>czerwony</seg>`.

While, in many cases, information on boundedness could be recovered from the morphosyntactic description of the given segment, this does not resolve the issue because, firstly, a recourse to morphosyntactic annotation layer in order to recover information lost in the segmentation layer is methodologically flawed (in some cases, it is perfectly imaginable that a text is only accompanied by the segmentation layer of annotation and nothing else), and, secondly, morphosyntactic identity will not resolve all such cases. Consider the example of *żółto-czerwony* "yellow and red": the segment *czerwony* here is bound, but both graphically and morphosyntactically identical to the frequent free-standing segment *czerwony* "red".

In order to accommodate such cases, we have defined an additional attribute of the <seg> element, @nkjp:nps, where "nkjp:" is the non-TEI namespace prefix, while "nps" stands for "no preceding space" and its default value is "false". Naturally, this attribute solves issues specific to Polish and similar languages. It can be generalized and become something like @bound={"right", "left", "both"}, and in this shape, get incorporated into the TEI Guidelines.

4.2 Structural disjunction between alternative segmentations

One strategy to handle alternative segmentations, where the choice is between a single segment of

[3]Note that here, string-range() is an XPointer **scheme** defined by the TEI. It is not to be confused with the string-range() **function** of the XPointer xpointer() scheme, defined by the W3C permanent working draft at http://www.w3.org/TR/xptr-xpointer/.

the form <seg>*New York*</seg> and a sequence of two separate segments, <seg>*New*</seg> and <seg>*York*</seg>, is to perform radical segmentation (always segment *New* and *York* separately) and provide an extra layer of alternative segmentation that may link the two parts of the name into a single unit. This is what we do in the creation of the annotation level of syntactic words that may, e.g., need to link the three segments of *żółto-czerwony* above into a single unit, because this is how they function in the syntactic representation.

In some cases, however, radical segmentation may create false or misleading representations, and Polish again provides numerous relevant examples. Sometimes bound segments, such as the person-number clitics illustrated in (2) above, are homophonous with parts of words.

(5) a. miał|em vs. miałem
had.prt|1sg fines.instr.sg
b. czy|m vs. czym
whether|1sg what.instr
c. gar|ście vs. garście
pot.acc|2pl fistful.nom.pl

One may attempt to defend radical segmentation for case (a) on the not-so-innocent assumption that segmenting tools might sometimes reach inside morphological complexes and separate affixes from stems, rather than clitics from their hosts. However, examples (b) and (c) show that this is not a feasible approach here: the Instrumental *czym* in (b) is monomorphemic, and the segmentation of *garście* "fistfuls" into *gar-* and *-ście* is likewise false, because the putative segment division would fall inside the root *garść*.

Thus, radical segmentation is not an available strategy in the case at hand. What we need instead is a way to express the disjunction between a sequence such as <seg>*miał*</seg> <seg>*em*</seg> (cf. (5a)) on the one hand, and the single segment <seg>*miałem*</seg> on the other. It turns out that the TEI has no way of expressing this kind of relationship structurally.

The TEI Guidelines offer the element <choice>, but it can only express disjunction between competing segments, and never between sequences thereof. The Guidelines also offer two non-structural methods of encoding disjunction. The first uses the element <join> (which is an ID-based equivalent of a bracket – it points to the segments that are to be virtually joined) and the element <alt> (which points at encoding alternatives). The other utilizes the @exclude attribute, which, placed in one segment, points at elements that are to be ignored if the segment at hand is valid (the excluded elements, in turn, point back at the excluding segment).

Recall that the intended size of the corpus is one billion segments. Tools that process corpora of this size should not be forced to backtrack or look forward to see what forms a sequence and what the alternative to this sequence is. Instead, we need a simple structural statement of disjunction between sequences. The solution used by the NCP consists in (i) adding an element meant to provide a semantically neutral bracket (<nkjp:paren>) and (ii) including <nkjp:paren> in the content model of <choice>. Note that this representation can be readily converted into the pivot format of the LAF:

(6)
```
<choice>
  <seg>miałem</seg>
  <nkjp:paren>
      <seg>miał</seg>
      <seg nkjp:nps="true">em</seg>
  </nkjp:paren>
</choice>
```

5 Conclusion

We have presented the TEI-P5-XML architecture of the National Corpus of Polish and identified some weak points of the TEI-based stand-off approach: the impossibility of keeping the primary text unannotated in the XInclude system, the loss of information on segment-boundedness, and the absence of a structural statement of disjunction between sequences of segments (this last issue is also due to the lack, among the numerous detailed markup options provided by the TEI, of a semantically neutral bracket-like element whose only role would be to embrace sequences of elements).

We are grateful to the two anonymous LAW-09 reviewers for their helpful comments.

References

Ide, N. and L. Romary. (2007). Towards International Standards for Language Resources. In Dybkjaer, L., Hemsen, H., Minker, W. (eds.), Evaluation of Text and Speech Systems, Springer, 263-84.

Przepiórkowski, A., R. L. Górski, B. Lewandowska-Tomaszczyk and M. Łaziński. (2008). Towards the National Corpus of Polish. In the proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008), Marrakesh, Morocco.

TEI Consortium, eds. 2007. TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 1.2.0. Last updated on February 1st 2009. TEI Consortium.