# The Pros and Cons of Stand-off Annotation:
# IPI PAN Corpus Design

Piotr Bański, University of Warsaw
bansp@ipipan.waw.pl

The problem outlined in the present paper presented itself in the planning stage of a project aiming at constructing the IPI PAN corpus of written Polish. The project is funded by the State Committee for Scientific Research grant no. 7 T11C 043 20. The IPI PAN corpus is going to contain at least 75-100 million words, and is going to be constructed with many diverse applications in mind, although these will primarily be related to language engineering. It is also going to be annotated structurally and morphosyntactically according to the suggestions laid out in the Corpus Encoding Standard Guidelines (Ide *et al.* 1996). The corpus will contain several sub-corpora divided according to the genre of the texts that make them up (e.g., literary texts, dialogue transcripts, etc.), as well as a balanced reference subcorpus that should be representative of modern standard Polish, and a hand-verified subcorpus designed for the purpose of training the morphosyntactic tagger. In what follows, we briefly report on a class of design problems related to the use of the so-called stand-off morphosyntactic and structural annotation, advocated by the Corpus Encoding Standard.

## 1. Encoding scheme

In several respects, the IPI PAN corpus resembles the American National Corpus (ANC). Most importantly, both these corpora are being created according to the XCES guidelines for corpus encoding. XCES (the XML version of the Corpus Encoding Standard, see Ide *et al.* 2000 and http://www.cs.vassar.edu/XCES), provides both a data architecture suitable for linguistic corpora and an encoding standard that expresses this architecture. Furthermore, it describes in detail the subsequent stages of conformance for corpus building, with the basic stage being the most cost-effective to achieve and at the same time suitable for basic NLP and general applications, and the final stage expressing the most detailed annotation information for specialized NLP applications.[1]

## 2. Gross corpus structure

The following are the two possible expansions of the `<cesCorpus>` element, which is the main structural unit of text arrangement within the corpus:

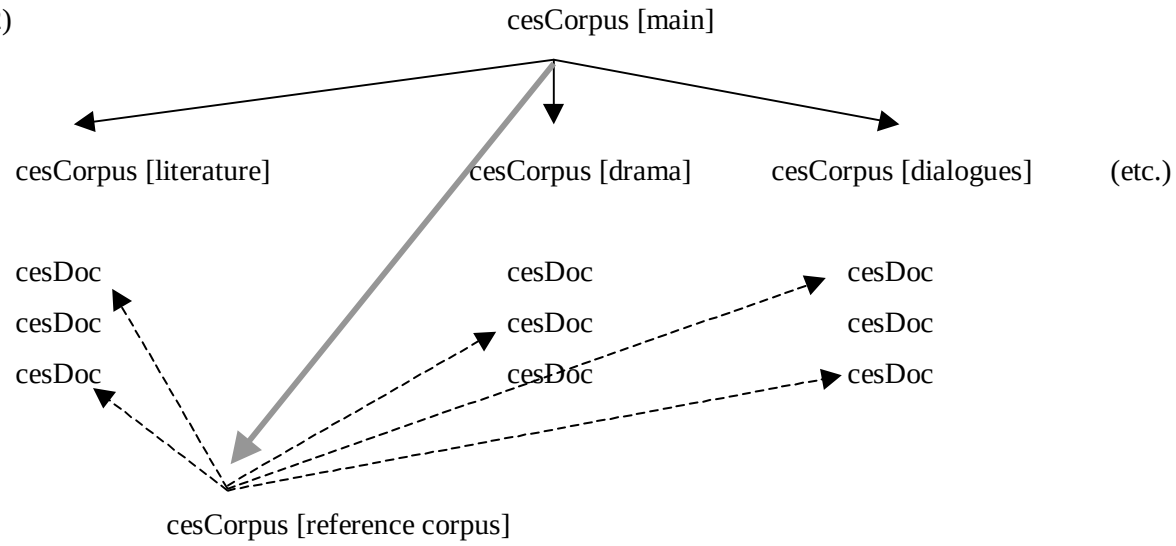
(1) a.
```
<cesCorpus>
    <cesHeader>
    </cesHeader>
    <cesDoc>          [one or more]
    </cesDoc>
</cesCorpus>
```

[1] For a survey of more technical issues involved in the creation of the IPI PAN corpus, see Bański (2001).

```
b.<cesCorpus>
      <cesHeader>
      </cesHeader>
      <cesCorpus>          [one or more]
          <cesHeader>
          </cesHeader>
          <cesDoc>         [one or more]
          </cesDoc>
      </cesCorpus>
  </cesCorpus>
```

The IPI PAN corpus uses only one-level-deep nesting of the `<cesCorpus>` elements, according to the primary divisions based on the genre of the texts included in it. Notice that because some subcorpora (notably the reference subcorpus for modern standard Polish) are meant to cut across the major division into literary genres, this system may not straightforwardly be used to describe all of them. Instead, a logically separate cesCorpus file will be needed for this purpose. This separate file will include the relevant documents from all the major subcorpora. The basic corpus architecture is sketched below.

(2)




Technically, nothing prevents the inclusion of the reference corpus as yet another subcorpus of the main <cesCorpus> element, as indicated above by the thick gray arrow. In fact, this is a welcome solution, in that the main corpus file should include all of the others. However, one should be aware that logically, the reference subcorpus belongs on a different plane.

**3. Stand-off annotation**

As mentioned above, all corpus texts will be annotated with morphosyntactic information. This is going to be an instance of so-called *stand-off* (*remote*) *annotation*, as advocated by the (X)CES. In this system, the annotation information is located in a separate file that indirectly references the main text,

annotated with gross structural information down to the level of the paragraph and made read-only. In this case, the hypertext links between the pieces of annotation and pieces of the main text express semantic information: they identify places in the original text where the given annotation should appear, in effect creating a kind of virtual markup.

### *3.1. Advantages of stand-off annotation*

Stand-off annotation has several advantages over the traditional method whereby all markup is stored together with the original text. These advantages are summarized below (see e.g. <http://www.cs.vassar.edu/CES/CES1-5.html#ToCOview> for more details).

- The original is kept as a read-only document, containing gross structural markup only. This means that there is no risk of accidental data corruption as new annotation is added.
- New annotation documents can be created and linked to the original at any time.
- There can be e.g. multiple morphosyntactic annotation documents, depending on the particular theory of morphology and syntax applied, and on the analyzer used. This is useful for cross-theory comparisons, as well as for judging the effectiveness of various analyzers.
- Simple searches should be faster, as there is less text to process in queries that do not use morphosyntactic criteria.
- The original documents from monolingual corpora may be reused in the creation of parallel or comparable corpora.[2]
- The problem of overlapping hierarchies is avoided, because the two (or more) hierarchies in question will be kept in separate files. This problem manifests itself in the case of the so-called bracketing paradoxes, or the division into verses and sentences in poetry, or quotations and sentences in literary texts, transcriptions of multi-party dialogues, etc.[3]
- In connection with the preceding issue, remote annotation makes it possible to easily create multiple views of the document, depending on what features of the text are relevant from the point of view of the user. This becomes trivial if the XML version of the CES is used, as an XML-encoded corpus may be easily transformed by XSLT scripts into e.g. the HTML or (La)TeX formats and, when necessary, rendered by XSL/CSS stylesheets; there exist numerous engines capable of effecting such transformations, and the WWW interface may be created with e.g. Perl or Python CGI scripts using various XML-handling modules.[4]

---

[2] See <http://www.ilc.pi.cnr.it/EAGLES96/corpustyp/corpustyp.html> for the characterization of these two types of multi-lingual corpora.

[3] See <http://www.cs.vassar.edu/CES/CES1.Annex10.html> for more discussion on resolving the problem of overlapping hierarchies within the CES.

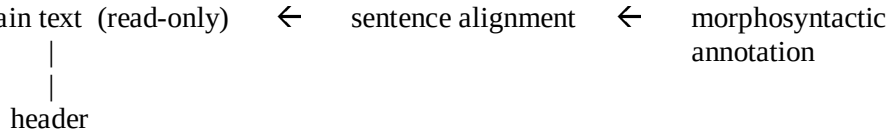
[4] Because of space limitations, we do not adduce concrete examples of the XML implementation of various encoding strategies. See e.g. Ide *et al.* (2000) for illustrations of XSLT scripts, and XML encoding using the XPointer and XML Schema techniques used in corpus management.

- As Ide (1998) remarks, the stand-off annotation architecture allows for better control of the base documents while allowing for simultaneous free distribution of the markup. This is a nontrivial issue for corpora bound by various non-disclosure licenses.

### *3.2. Implementation of stand-off annotation in the IPI PAN corpus*

The gross file structure for a single text in the corpus is presented below:

(3)




The header of the main text is located in a separate writeable file, included into the main read-only text as an external XML entity. This is because modifications to the header may need to be made more frequently than those made in the main text, the latter being mainly corrections of typographical errors, if any. In this way, the read-only status of the main text protects it from accidental corruption while the header is available for editing.

For the IPI PAN corpus, two kinds of stand-off annotation documents are initially used. The first kind contains sentence segmentation for the given text. It refers to the base text via one-way links. The other kind of remote markup documents contain morphosyntactic annotation. This kind of annotation does not reference the base text but rather the document with `<s>`-alignment annotation, as illustrated above. The sentence alignment document has the `version` attribute of the `<cesAna>` element (the root element of the XML tree containing information about grammatical analysis of the given text, see <http://www.cs.vassar.edu/CES/CES1-5.html>) set to "sent", whereas the morphosyntactic annotation document is set to "tok lex disamb", to signal that it deals with text tokens and their possibly multiple interpretations (included in `<lex>` elements), and also the disambiguated forms.

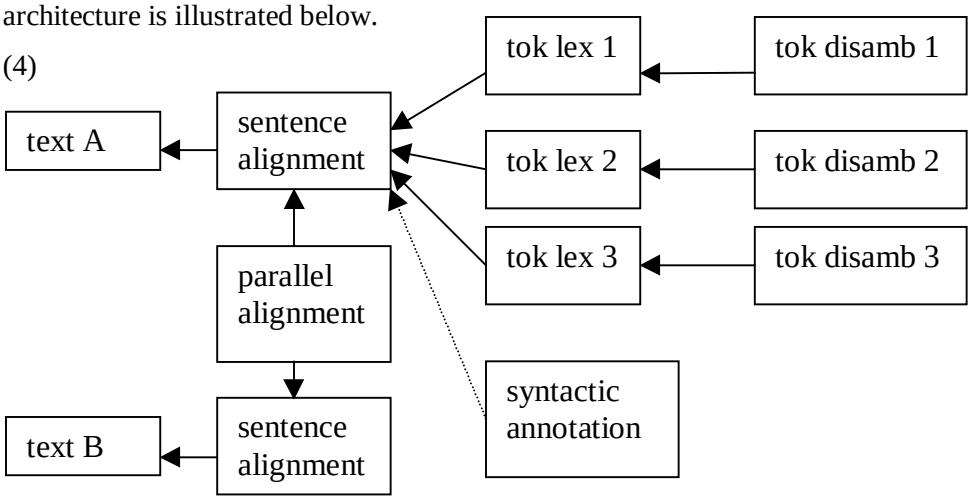
In this way, the sentence alignment documents in the middle act as the base for various possible morphosyntactic annotation documents, as well as documents containing e.g. syntactic or semantic annotation, which may be added at a later time.[5]

Morphosyntactic annotation includes POS markup and basic morphological information. It also identifies the lemmata. For the detailed specification of the morphosyntactic tagset, see Woliński & Przepiórkowski (2001). In the early stages of the project, several such files may be used to reference a single text, for the purpose of comparing the accuracy of various morphosyntactic analysers and disambiguators. The architecture outlined above makes it possible to create more than a

[5] As discussed by Woliński & Przepiórkowski (2001), disambiguation in the basic version of the IPI PAN corpus morphosyntactic annotation is restricted to the domain of the <s>, which makes the sentence-alignment document the ideal target for remote morphosyntactic markup.

single type of morphosyntactic annotation document: instead of a single “tok lex disamb” document, some documents may be marked as “tok lex” (i.e., containing information on tokenization and the output of morphological analysers) and others as “tok disamb” (containing the disambiguated tokens output by the disambiguator). Such a division of data may be useful for the purpose of comparing the efficiency and precision of various morphological analysers and disambiguators. This extended architecture is illustrated below.

(4)




Many other kinds of annotation documents may be added in an analogical fashion, containing for example (where appropriate) discourse or semantic annotation.

The above picture does not claim to depict the only logically possible arrangement: the syntactic annotation document may refer to the sentence alignment document, as indicated above, but it could just as well refer to the documents containing disambiguated output of the tagger. In the former case, the information from the “tok disamb” documents and syntactic annotation documents has to be merged indirectly, via suitable XSLT scripts. In the latter case, syntactic annotation refererences the disambiguated tokens by their IDs. Similarly, the “tok disamb” documents need not reference the “tok lex” ones – they may reference the sentence-alignment document in the same manner as the latter. Each such logical arrangement requires a conscientious design decision, and some such decisions still await to be confirmed by experience. One factor that should not be ignored is the ease of maintenance of the corpus – for example, if both the “tok lex” and “tok disamb” documents index the sentence-alignment document in the same manner, then some changes performed in the main text (such as changes in the number of tokens within each <s> unit) will need to be compensated for twice as many times than in cases when only the “tok lex” documents count words within the sentence-alignment documents and the “tok disamb” documents merely refer to the ID numbers of the tokens contained in the “tok lex” documents.

The kind of architecture illustrated in (4) above will also allow for better reuse of the corpus, if in the future some of the texts will be used as parts of parallel or comparable (sub)corpora. In such cases, the relevant parallel alignment documents will contain two-way links addressing the appropriate sentence alignment documents (see <http://www.cs.vassar.edu/CES/CES1-5.html#ToCalign>). Notice

also that parallel alignment documents need not come into play only in multilingual corpora other than the IPI PAN corpus described here. It may be useful for e.g. translation studies to be able to compare two or more Polish translations of the same foreign text, all of which can be proper parts of the IPI PAN Corpus. Examples are not hard to find: a classic example is various translations of the Bible, another more lightweight example is several Polish translations of "The Lord of the Rings" by J. R. R. Tolkien, which are already the subject of fierce Internet debates of fantasy fans.[6] Turning to matters more sublime, we may also mention the numerous translations of e.g. Shakespeare's plays, among many others.

### *3.3. Tokenization*

At first glance, it would be tempting to include the tokenized text in the same file as that containing sentence-level segmentation: in this manner, this could be the only file using XPointer mechanisms to index the original document on a character-by-character basis. Because both `<s>` and `<tok>` elements possess ID attributes, all the other annotation documents could merely refer to their ID values, which would, among other things, make those other documents smaller. However, in order to satisfy all possible morphosyntactic analyzers and disambiguators, tokenization would have to be very radical. Divisions based on spaces, as in the case of e.g. *po prostu* 'simply', which could in many cases be treated as a single token, are not enough. A decision should be made whether to divide e.g. *żółto-zielony* 'yellow intermingled with green' into two separate tokens, and if so, whether to divide *żółtozielony* 'green with a tinge of yellow' into separate tokens as well. The same question can be asked about the so-called movable or clitic auxiliaries, as in *szybkośmy* 'fast+1PL', where such a division makes a lot of sense vs. *zrobiliśmy* 'do-PARTICIPLE+1PL; we did', where the clitic *śmy* '1PL' can be analyzed as an inflectional ending on the verb (see Bański 2000 for extensive discussion). Finally, examples such as *stodwudziestopięciokrotny* 'one-hundred-twenty-five-fold' might be radically analyzed as sequences of *sto* 'hundred' + *dwudziesto* 'twenty' + *pięcio* 'five' + *-krotny* '-fold', or as sequences of two tokens, the numeral and the bound stem *-krotny*.[7] By locating tokenization information in the given annotation file, we let the particular morphosyntactic analyzer and/or disambiguator impose their own requirements on the degree to which it is necessary to divide text chunks.

[6] See e.g. <http://www.gazeta.pl/alfa/home.jsp?dzial=0511&forum=139> or the pl.rec.fantastyka.sf-f newsgroup as the starting points. See also <http://www.republika.pl/tlumok/lozins1.htm> for a partial (ambiguity intended) comparison of two of the available four Polish translations of the book.

[7] I am grateful to Adam Przepiórkowski for a discussion on the issue of radical tokenization.

### *3.4. Problems concerning stand-off annotation*

This section concentrates on problems posed by the kind of architecture adopted for the IPI PAN corpus. We have already mentioned some of them: there is no single way to follow the existing standards. The complete new version of the XCES is still not released to the public at the time of writing of the present paper, and there are few hints that can be found in the already existing corpora (the MULTEXT-EAST corpora, for example, do not instantiate a pure version of the CES, and there is no publicly available corpus that we know of that would fully instantiate the XCES system, even in its early version). Some part of the development path will surely have to be laid out by trial and error – this, however, is part of the excitement that made the development team decide to take up the challenge in the first place.

As pointed out by Martin Wynne (personal communication), a disadvantage of remote annotation shows up when it uses character-by-character indexing and when it is necessary to correct e.g. some typographical errors in the original. This may in most cases force re-indexing of the parts of all the remote documents which address text fragments within the scope of the element that encloses the corrected text. We accept this as partial cost of the kind of robust corpus structure described here. Care will be taken to minimize this kind of problems by using specially designed (re)indexing tools which will identify the extent of the corrections needed to be performed in stand-off annotation documents.

Other problems concern complex searches made, for example, according to lexical, structural, and morphosyntactic criteria at the same time. At first glance, such searches require access to several files at once. However, with the power of the XML Framework behind the XCES system, the issue becomes trivial: a more *ad hoc* way to perform such searches easily is to use XSLT scripts to merge the relevant files containing the required kinds of information into a single file, and to perform the search on this resulting file. A more intricate way is to make use of the existing standards for corpus interchange and maintenance. One notable standard is the ATLAS system based on the so-called Annotation Graph model endorsed by Steven Bird and Mark Liberman of the Linguistic Data Consortium (http://www.ldc.org/, see Bird & Liberman 2001). Another serious emerging standard is that built on the basis of the XCES itself, as well as ATLAS and other existing standards. This is the GMT (Generic Mapping Tool) model proposed by Nancy Ide (Vassar College) and Laurent Romary (LORIA/INRIA). This model currently serves as the starting point for the work of the ISO/TC 37/SC 4 Language Resource Management committee, see Ide & Romary (2001). Both the models come equipped with transducers from various XML (and in fact also non-XML) formats into their native formats, as well as various tools that operate on those native formats and that can be used for corpus management. With this sort of tools at our disposal, the problem of complex searches, and in general, the problem of corpus maintenance and interchange, disappears.

A final, minor design issue concerns the fact that the CES recommends that pieces of the original text be included in the remote annotation documents. This was possibly at least partially

caused by the need to make hand validation easier or to make searches over, or display of such data, more convenient. However, given that by means of appropriate XSLT scripts the textual data and the annotation can easily be put together in any form, whether plain text, HTML or other, there seems to be no need to extend the corpus size by including pieces of the original text elsewhere. The more so that this raises potential problems when it comes to e.g. correcting typographical errors in the original text and making sure that these changes are repeated in every document that happens to include copies of the original. It is much safer to store the textual data in one place only, namely the original read-only document, and to use extended pointers to reference these data from elsewhere.

## 4. Summary and conclusion

We have briefly reviewed the issue of implementing a stand-off annotation system for a large text corpus. We have looked at some of the possible architectural solutions, tokenization issues, as well as more general problems regarding the adoption of the XCES framework. We conclude that despite the apparent and real hardships, the idea of spreading specified parts of structural and morphosyntactic information over several files is an exciting and fruitful way to encode a modern corpus with an eye towards its future extensions, reusability, and interchange.

## References


Bański, Piotr (2000). Morphological and phonological analysis of auxiliary clitics in Polish and English. PhD. Dissertation, University of Warsaw.

Bański, Piotr (2001). The proposed encoding scheme for the IPI PAN corpus. Technical report, Polish Academy of Sciences.

Bird, Steven, and Mark Liberman (2001). "A formal framework for linguistic annotation". *Speech Communication*, 33, 23-60.

Ide, Nancy (1998). "Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora." *Proceedings of the First International Language Resources and Evaluation Conference*, Granada, Spain, 463-70.

Ide, Nancy, Priest-Dorman, Greg, and Jean Véronis (1996). Corpus Encoding Standard. Available at <http://www.cs.vassar.edu/CES/>.

Ide, Nancy, Bonhomme, Patrice, and Laurent Romary (2000). XCES: An XML-based Standard for Linguistic Corpora.. *Proceedings of the Second Language Resources and Evaluation Conference (LREC)*, Athens, Greece, 825-30.

Ide, Nancy, and Laurent Romary (2001). Standards for Language Resources. Department of Computer Science, Vassar College and Equipe Langue et Dialogue, LORIA/INRIA.

Woliński, Marcin and Adam Przepiórkowski (2001). Projekt sposobu morfosyntaktycznego anotowania korpusu języka polskiego. Technical report, Polish Academy of Sciences.