

Piotr Bański

**The proposed encoding scheme for
the IPI PAN corpus^{*}
(7 T11C 043 20)**

Nr

Warszawa, grudzień 2001

Streszczenie

Praca opisuje standard opisu strukturalnego i morfosyntaktycznego znany jako CES, czyli Corpus Encoding Standard, skupiając się na jego nowoczesnej implementacji w języku XML. Standard ten będzie użyty do oznaczenia dużego korpusu tekstów języka polskiego przygotowanego w Instytucie Podstaw Informatyki PAN w Warszawie. Autor omawia przyczyny wyboru tego właśnie standardu, a w szczególności jego wersji zdefiniowanej w języku XML. Praca kończy się opisem praktycznych czynności jakie będą wykonane w ramach konstruowania korpusu.

Słowa kluczowe: anotacja, Corpus Encoding Standard, inżynieria języka, korpus tekstowy, XML

The present report describes the encoding scheme used for the purpose of creating a large morphosyntactically annotated corpus of Polish (henceforth the IPI PAN corpus), funded by the State Committee for Scientific Research grant no. 7 T11C 043 20. The 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 subcorpus designed for the purpose of training the morphosyntactic tagger.

The report begins with a brief overview of the role that SGML (Standard Generalized Markup Language) and XML (eXtensible Markup Language) play in the realm of corpus linguistics. Section 2 looks at the CES (Corpus Encoding Standard), the encoding standard forming the basis for the annotation of the IPI PAN corpus. Section 3 reports on the major components of the so-called XML Framework, i.e., the XML-related standards designed for text indexing, manipulation, and description. It also points out the shift of focus that the adoption of the XML Framework effected in the field of language resources, and specifically, language corpora. Section 4 describes the overall structure of the IPI PAN corpus, and discusses the way in which encoding of this corpus should proceed. Annex 1 lists the

^{*} I would like to express my thanks to Adam Przepiórkowski for his comments on a previous version of this report. I am also very grateful to Nancy Ide for a discussion on the new version of the XCES system as well as for her help in

abbreviations and acronyms used in the text, together with the relevant URLs. Annexes 2 and 3 contain the xcesDoc and xcesAna DTDs, respectively.

1. Metalanguage standards applied in text encoding

Among the metalanguage level standards able to handle linear signals indexed by intervals, such as speech or text, at present only two enjoy serious interest and popularity within the field: SGML (Standard Generalized Markup Language) and XML (eXtensible Markup Language). That is why already at the outset, we restrict ourselves to considering only SGML-based and SGML-derived encoding systems, in which non-procedural, *descriptive* markup schemes can be expressed.¹

SGML, while made largely obsolete by XML, is still widely used in document preparation (in library cataloguing systems, help/manual files on many computer systems, etc.) and specifically, corpus encoding (for example in the British National Corpus or the Polish National Corpus, among others). It should also be borne in mind that HTML (HyperText Markup Language), which revolutionized the Internet and which for many Internet users *defines* ‘the Web’, is an application of SGML. XML is a simplified version of SGML and therefore it is less powerful than its predecessor, which among other things leads to fewer problems concerning automatic validation. On the other hand, XML overcomes problems caused by its smaller expressive power thanks to the existence of its auxiliary specifications, such as XML Namespaces, XPointer or XLink, to name but a few (see <http://www.w3.org/>, and section 3). It is also possible to express constraints on XML markup by means of the so-called XML Schema language, advocated by the World Wide Web Consortium (W3C) or its various non-W3C-bound alternatives, such as the Schematron or RELAX NG schema languages, among others. This not only largely eliminates the need for Document Type Definitions (DTDs) but also makes it possible to express conditional constraints on the content, occurrence, or attributes of an element, based for example on the syntactic context in which this element occurs.^{2,3} Unlike for SGML, a vast amount of freely available (and most often, Open Source, hence easily customizable) software packages exist for XML. This adds to the cost-effectiveness of projects utilizing XML. Additionally, because HTML has been redefined as an XML application under the name of XHTML, XML is on its way to win over SGML as the everyday all-purpose Web standard.

The two SGML-based encoding systems considered here are the Text Encoding Initiative Guidelines (TEI), version P3 and the Corpus Encoding Standard (CES). Both of them are in the process of developing XML incarnations, referred to as TEI P4, and XCES, respectively.

locating the relevant references.

¹ Descriptive markup identifies the logical and structural parts of a text, with no attempt made to influence their rendering by user agents, i.e., particular programs interpreting the document language.

² Document Type Definitions (DTDs) are a legacy of SGML used to define various properties of XML documents, such as their syntax, default values of attributes, obligatory or optional elements, their basic data content, macro functionality (via parameter entities), as well as sets of replaceable elements of various kinds (entity references and character references). DTDs require a specific, non-SGML syntax. While they are not obligatory, they ensure control over the document content and are necessary for interchange (understood as tag-sharing, see also section 3.2).

³ To be precise, schema languages do not replace DTDs in their entirety but rather provide an alternative and XML-compliant way of expressing the *modeling* aspects of DTDs, i.e. the structural constraints. See e.g. <http://lists.xml.org/archives/xml-dev/200106/msg00579.html> (Eric van der Vlist’s post to the xml-dev list) for more discussion of this issue with regard to the XML Schema language. See <http://www.oasis-open.org/committees/relaxng/tutorial-20011203.html#IDAGSZR> for a comparison between RELAX NG and XML DTDs. See also section 3 for some further details.

Version P3 of the TEI Guidelines was released in 1994 and since then it has achieved enormous success as a text encoding and interchange standard (see the page at <http://www.tei-c.org/Applications/index.html>, listing major projects based on this specification). The CES was created in 1996 as a TEI-based encoding scheme designed specifically for the purpose of providing a data architecture for linguistic corpora together with a set of encoding principles expressing particular levels of that architecture. These principles range from the simplest and the most cost-effective scheme (constituting a minimal specification that linguistic corpora must comply with in order to be considered suitable for basic language-engineering applications) to the most refined scheme fully reflecting all the properties of texts necessary for specialized NLP applications. The XML version of this scheme (XCES) has been adopted for the purpose of encoding the IPI PAN corpus.

On the 1st of August, 2001, the first official release of version P4 was announced. The major innovation of the TEI P4 is that unlike P3, this release is 'XML-ized', i.e., its architecture allows for handling of both SGML and XML, depending on how it is configured.⁴ The final version is expected at the beginning of the year 2002. The full release of the TEI P4 may influence the project described here, marking the migration from free-standing XCES DTDs to XCES applied as a set of extensions to the TEI P4.⁵

The decision concerning whether to recreate the existing XCES DTDs as extension files to the TEI P4 DTDs or whether to switch to XML Schema (or RELAX NG) constraints will remain a challenging question for the future of the IPI PAN corpus. For the moment, both alternatives present certain advantages and drawbacks, and given that both still await their completion or wider acceptance within the corpus community, the conservative decision to remain with XCES DTDs at least at this early stage of the project, appears to be the safest one.

2. Corpus Encoding Standard

The Corpus Encoding Standard was created in co-operation between the EAGLES (Expert Advisory Group on Language Engineering Standards) project, the MULTEXT (Multi-lingual Texts and Tools for Western European Languages) project, as well as other projects undertaken jointly by the Vassar College and Equipe Langue et Dialogue, LORIA/CNRS. It has been used in the PAROLE, TIPSTER, and MULTEXT-EAST projects, among others.⁶ Currently, its XML version is the standard adopted for encoding the American National Corpus (ANC); to some extent, it has also been adopted in the MATE project.

The CES is an application of SGML created according to the TEI guidelines and specifically designed to constitute a set of encoding standards for language corpora, with a view to their maximal usefulness in language engineering. It also provides encoding specifications for linguistic annotation, together with a data architecture for linguistic corpora.

2.1. Major design principles behind the CES

The CES has been designed with three main purposes in mind (Ide 1998):

⁴ The so-called TEI-Lite standard was the first version of the TEI with an XML DTD; however, this version in itself is not suitable for corpus encoding, as it lacks the appropriate functionality, being a drastically impoverished version of the full TEI standard.

⁵ Version P4 is going to be equipped with an experimental RELAX NG schema form (Sebastian Rahtz, email message to the TEI-L mailing list).

- to reflect best practice and consensus within the corpus community,
- to be maximally usable and reusable,
- to be cost-effective.

As far as the first issue is concerned, the CES, itself developed on the basis of a corpus-building project, also drew on the TEI experience, and on the input from many other corpus-encoding projects. The success of the above-mentioned projects that adopted the CES architecture, mentioned earlier in this section, testifies further to its value.

The second of the above-mentioned issues, naturally connected with the first, refers to the maximal suitability for many language-engineering purposes, as well as – in connection with the issue of cost-effectiveness – also to the reuse of CES-based corpora for purposes other than originally intended, and the reuse of the tools created to process such corpora. This becomes especially relevant and true for the XML version of the CES.

As for the third issue – cost-effectiveness – understood as distinct from the issue of (re)usability, attention should be drawn to the ways in which texts are currently obtained. For the most part, modern corpora are built from *legacy data*, i.e., texts already existing in electronic form, and most often containing some kind of encoding, standardly referred to as *foreign markup*. To make such texts suitable for inclusion into CES-compliant corpora, the foreign markup has to be replaced with CES-based markup. This means that some of the foreign markup, such as column divisions or ‘hard’ page breaks not essential for the interpretation of the text, have to be removed. Other bits of foreign markup, such as various kinds of *rendition information* concerning the shape of the text (for example passages in bold or italicized font) which will not be visible in plain text format, also have to be suitably re-encoded. This kind of CES-compliant markup, if it is generic enough (e.g., signaling that the given text span is highlighted, without specifying what this highlighting denotes), may be created in an automated fashion. On the other hand, marking up the specific information on whether the given type of text highlighting in the given context marks a foreign term or emphasis, etc. requires human intervention and is therefore costly. As mentioned above, the CES is in fact a set of encoding standards. This means that the CES guidelines provide a series of encoding schemes, with an increasing degree of refinement, so that the least detailed scheme may be used for automated and cost-effective basic encoding of raw legacy texts, and the most detailed scheme is meant to be maximally useful for the purpose of broadly defined language engineering (specifying what the original rendition information stood for, identifying names, dates and abbreviations, which should be treated in special way by parsers, etc.).

The original, SGML-based CES can be thought of as an application of the TEI in that it was created by adding the appropriate set of DTD fragments as so-called extension files to the main TEI DTD. Because the CES targets language corpora and their applications in language engineering, on the one hand, it limits the power of the original TEI scheme and on the other, it adds extra features necessary for the proper handling of corpora.

One reason to limit the power of the TEI scheme is that, being an interchange standard, the TEI is by design extremely general – as an interchange standard, it must be general enough for any local standard to be transducible into it without any loss of information. Partially as a consequence of that, the TEI usually offers several element sets or even mechanisms (in the case of feature structures) to handle a single encoding problem. Because the CES is created largely as a local processing standard, it takes into consideration the processing cost of the various encoding solutions offered by the TEI. We return to this issue in section 2.3. Before that, though, we take a look at the most important features and advantages of the CES.

⁶ See <<http://www.cs.vassar.edu/CES/CES-P.html>> for a list of CES-based projects.

2.2. Advantages of the CES

One of the main advantages of the CES is that it tailors down the robustness of the TEI to the task of managing corpora. This does not merely mean choosing a set of tags appropriate for all corpora and corpora alone. One of the crucial factors in the creation of the CES scheme was the efficiency with which the resulting annotated corpus can be processed.

On the one hand, the TEI overgenerates in some respects, in that it offers far too many ways of handling a single encoding problem. On the other, it is too poor – it offers too few options for the needs of (complex) corpora. The CES adds some elements and modifies some of the existing ones. A trivial example of that is that the `<CesCorpus>` tag may be recursively nested – as opposed to the `<TeiCorpus.2>` tag – thus allowing corpus creators to introduce subcorpora. We return to both these issues in the following section.

As already mentioned, the CES is in fact a set of standards, ranging from the simplest type of annotation, achieved in most cases almost cost-freely, to complex annotation that needs human intervention on the one hand, but on the other, it is very well suited for NLP applications. Hence, it is possible to start off small and cheap at the beginning, and – in time – to refine the annotation, when more financial resources and manpower are available. What is important is that the path of this so-called *up-translation* is already laid out by the CES guidelines.

Furthermore, the CES is specifically designed to handle remote (stand-off) annotation, and given the existing robust XML-based mechanisms (XPath, XPointer, XLink, XSL(T), XML Schema, etc., together with a range of freely accessible parsers/drivers handling them), processing of such ‘distributed’ corpora is currently unproblematic. We return to the issue of remote annotation in section 2.4.

2.3. Consistency and precision

The overall goal of the CES is “identification of a minimal encoding level that corpora must achieve to be considered standardized in terms of descriptive representation . . . as well as general architecture.” (Ide & Véronis 1993) These two goals pertain, respectively, to the proper markup of structural and typographic information within the corpus, and to its maximal suitability as a linguistic resource.

There is a certain degree of tension between the two aims, if ‘descriptive representation’ is taken to extend to the questions of data interchange: for a markup scheme to be suitable for the purpose of data interchange, it must be extremely general and flexible, so as to cover and absorb all the possible existing annotation systems. This is what the TEI is aimed to be. However, a general scheme such as the TEI allows e.g. for multiple encoding options for the same encoding problems and consequently, for a large degree of variation in encoding styles of even a single corpus.

One problem of indeterminacy concerns the choice of for example `<div>` vs. `<div1>`, `<div2>`, etc. tags, to mark text divisions. or `<s>` vs. `<seg type="s">`). Another concerns the information about the type of some tags given either in the form of an element name or an attribute value (cf. `<s>` vs. `<seg type="s">`). See Ide & Véronis (1993) for more discussion on these issues.

Below, yet another such case is illustrated, concerning the fact that a number of tags can at the same time nest others and be nested by them, which leads to potential confusion both within the corpus and even within a single document. Such confusion – as ever – translates straightforwardly into processing problems. As shown below, within

the TEI scheme, three elements are needed to mark up *tho*, the colloquial form of the word *though*, which is additionally emphasized. These three elements can be nested in any order, which creates an enormous potential for confusion both on the part of human encoders and the processing applications.

- (1) a. `<w><emph><orig reg="though">tho</orig></emph></w>`
- b. `<w><orig reg="though"><emph>tho</emph></orig></w>`
- c. `<emph><w><orig reg="though">tho</orig></w></emph>`
- d. `<emph><orig reg="though"><w>tho</w></orig></emph>`
- e. `<orig reg="though"><emph><w>tho</w></emph></orig>`
- f. `<orig reg="though"><emph><w>tho</w></emph></orig>`

According to the CES guidelines, all the original text must constitute tag content, and no tag content should contain anything else but original text. More specifically, information about properties of text, or notes and comments that do not constitute part of the original, should be restricted to occurring as attribute values only. This is an issue related straightforwardly to processing performance. As also illustrated in (1) above, the TEI allows for inconsistency with regard to such matters (Ide and Véronis 1993). In this case, what is at stake is inconsistency with regard to where the `rend` attribute can be used: it can be used on e.g. the `<name>` tag, but it cannot be used on the `<w>` tag, hence the need to use another element, `<orig>`, indicating the original form. While seemingly small when occurring in separation, it is such inconsistencies that pose enormous processing problems if they grow in numbers. The design of the CES strives to prevent such inconsistencies by appropriately constraining attributes and attribute values.

In general, in order to achieve higher precision and validatability, the TEI content models are substantially simplified within the CES. The TEI features the concept of attribute and element class, which groups attributes and elements with similar function or appearing in the same structural contexts. Such classes often overlap and in some cases form a complex hierarchy. The CES adopts the idea of attribute/element class, but limits itself to a shallow class hierarchy with no overlaps.

Elsewhere, the CES extends the TEI specification in order to cover the area of corpus encoding more thoroughly. Thus, it adds elements designed to provide a precise approach to encoding morphosyntactic features, reflected in the (x)cesAna DTD. It also suggests a specific element and attribute semantics together with specific recommendations for the kind of positional morphosyntactic tagging that easily translates into other formats.

2.4. Advantages of stand-off annotation

The CES advocates the use of the so-called *remote markup* (also called *stand-off annotation*), whereby the original document receives only minimal structural markup down to the level of the paragraph. All other markup is stored in separate files, and references the original via one-way hyperlinks, whose function can in this case be regarded as semantic, unlike in HTML where they are used for navigational purposes. In other words, the result of such linkage is that remote annotation is virtually added to the appropriate locations in the read-only original.

In the SGML-based CES, as in the TEI, where this mechanism originated, the HyTime (Hypermedia/Time-based Structuring Language) formalism was used to handle such links. XML has at its disposal the XLink language, which was built upon HyTime and the TEI concept of extended links. We return to the auxiliary XML specifications in

section 3.

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.

- 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.⁷
- 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.⁸
- 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 CES is used, as an XML-encoded corpus may be easily transformed by XSLT into e.g. HTML or (La)TeX 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.

2.5. CES conformance levels

There are three major conformance levels defined by the CES specification, level 1 being the most permissive and the least costly to achieve. The major conditions for conformance to each level are listed below (see <http://www.cs.vassar.edu/CES/CES1-4.html> for more detailed discussion). At each step, the resulting annotation document must validate against the (x)cesDoc DTD.

- Level 1:
 - Markup for major document divisions down to the paragraph level is required,
 - the header lists all the encoding formats used in the document,
 - no foreign markup is present.

⁷ See <http://www.ilc.pi.cnr.it/EAGLES96/corpusyp/corpusyp.html> for the characterization of these two types of multi-lingual corpora.

⁸ See <http://www.cs.vassar.edu/CES/CES1.Annex10.html> for more discussion on resolving the problem of overlapping hierarchies within the CES.

- Level 2:
 - The requirements for level 1 conformance are satisfied,
 - the rendition information is represented by the <hi> tag with an appropriate value of the rend attribute,
 - if a sub-paragraph element is marked, every occurrence of that element must be marked,
 - special characters (e.g. ‘pound’, ‘m-dash’) are identified and replaced with the appropriate XML entity references,
 - quotation marks are replaced either by the appropriate XML entities or <quote> or <q> tags.

- Level 3:
 - The requirements for level 2 conformance are satisfied,
 - the rendition information at the sub-paragraph level is determined (the <hi> element marking ‘highlighted’ words is resolved to concrete uses, such as <foreign>, <term>, etc.).
 - the paragraph level elements are correctly marked (as lists, quotes, ‘normal’ paragraphs, etc.),
 - sub-paragraph elements such as abbreviations, numbers, names, foreign words or phrases are marked either as the relevant element content (<abbr>, <num>, etc.) or by means of user-defined morphosyntactic tags,
 - the <p> (paragraph), <s> (sentence), and <q> (inline quote) elements have to nest in exactly this order (this requirement is naturally lifted if e.g. <s>-segmentation is stored in a separate stand-off annotation file, as it is going to be in the IPI PAN corpus),
 - finally, the encoding for all elements (excluding morphosyntactic tagging) must be validated for a 10% sample of the text.

It is common to assume that each conformance level will anticipate the requirements of the next one. Hence, it is recommended that e.g. <hi> tags are used already at level 1 and are properly resolved already at level 2, and so on. Information about the conformance level is stored in the level attribute (with values 1, 2 or 3) of the <conformance> element of the <editorialDecl> element of the <encodingDesc> element of the header. If it is not supplied there, it has to appear within the <projectDesc> element of the <encodingDesc> element.

3. XCES: the XML-ized version of the CES

In this report, we are looking primarily at the XML version of the CES guidelines, referred to as the XCES. Unlike the CES, which is an SGML application created by means of extension files added to the TEI P3 main DTD, the XCES has been created solely as a set of three free-standing DTDs, constituting an XML translation of the CES DTDs. Thus, the tie with the TEI is in this case indirect. This is naturally because the XML version of TEI did not even exist at the moment when the XCES was created (the beta version of the XCES dates from 2000).

3.1. XML Framework: overview of components

As we have already mentioned, XML’s core specification is greatly simplified with respect to its ancestor, SGML. Its

added power comes from auxiliary specifications which, together with the core XML specification, are sometimes referred to as the XML Framework. Some of the components of this framework relevant for the tasks of corpus encoding and maintenance are briefly reviewed below.

- XPath (XML Path Language, <http://www.w3.org/TR/xpath>) is a component part of XPointer and XSLT. Its primary purpose is to address (or, in XSLT applications, match) parts of an XML document. Additionally, it provides functions for basic string and number manipulation.
- XPointer (XML Pointer Language, <http://www.w3.org/TR/xptr>) builds upon the XPath syntax to provide a facility for addressing points and ranges within an XML document that are not exhaustively contained within parts of the XML tree structure.
- XLink (XML Linking Language, <http://www.w3.org/TR/xlink>) together with XPointer is the basis for an efficient mechanism offering methods of linking resources going far beyond the original HTML one-way hyperlinks. Its use in corpora designed on the basis of the CES is twofold: one-way links are used for virtual (stand-off) segmental and morphosyntactic annotation, and (at least) two-way links are used to align parallel corpora. See Ide *et al.* (2000) for remarks on how the SGML/HyTime-based TEI system for referencing locations in remote documents can be redefined and simplified by means of the auxiliary XML specifications.
- XSLT, XSL Transformations (<http://www.w3.org/TR/xslt>) is an autonomous part of the XML Stylesheet (XSL, <http://www.w3.org/TR/xml-stylesheet>). It allows for restructuring of XML documents and transforming them into other formats, such as HTML, LaTeX or plain text. Both the TEI P4 and the XCES come with a suite of basic XSLT scripts for data manipulation and display.
- Schema languages. Several such languages exist, the most popular being XML Schema, RELAX NG, and Schematron. They make it possible to create much tighter constraints on the document structure than DTDs could. Unlike DTDs, they also support namespaces (see below), and make it possible to define new, enhanced data types as well as constrain them. This is useful for feature analysis, and eliminates the need for Feature System Declarations used by the TEI in order to constrain feature structures (TEI P4, ch. 27). See Ide *et al.* (2000) for more discussion on the advantages of schema languages over DTDs in corpus creation and maintenance.
- XML Namespaces (<http://www.w3.org/TR/REC-xml-names>) make it possible for parts of the given document to follow different document structure definitions with – this is as if the given part of the document in question obeyed a DTD different from the DTD pertaining to the rest of the document. Thus, if an XML-encoded corpus is enriched with e.g. Dublin Core metadata (see <http://dublincore.org/>), the addition will involve the inclusion of the attribute `xmlns:dc="http://purl.org/dc/elements/1.1/"` in the header, so that all descendants of the element in which this attribute appeared, if they are preceded by the prefix `dc:` (note that ‘dc’ is the second element of the `xmlns:dc` attribute), refer to the document content model and semantics defined by the Dublin Core Metadata Initiative rather than the content model and semantics defined by the XCES.⁹ By establishing the full element or attribute name (the so-called ‘fully qualified name’) as a `prefix:local-name` pair (where the ‘local name’ is the bare tag), and by tying prefixes to unique identifiers (such as web locations or e.g. ISBN numbers), the namespace mechanism guarantees the uniqueness of element/attribute names, and makes it possible for a single document instance to be composed of interleaved

tags coming from many different DTDs or schemas.

- RDF, Resource Description Framework, (<http://www.w3.org/RDF/>) is concerned with metadata exchange and interoperability. It provides machine-readable descriptions of Web resources and complements the syntax-oriented XML-based specifications listed above, being essentially a semantic component of the XML Framework. RDF provides means to facilitate or enable activities such as web resource discovery, web content rating, data cataloguing, and others (see the FAQ list at <http://www.w3.org/RDF/FAQ>). It breaks away from the strictly hierarchical nature of XML and makes it possible to define class hierarchies that cut across XML trees (see e.g. <http://www.w3.org/TR/rdf-nt/> for more details on the model-theoretic semantics of RDF).

The adoption of XML as the metalanguage will make it easy for the creators of the IPI PAN corpus to reuse the existing XML tools and strategies. Several new annotation standards are currently emerging, designed to handle the issues of corpus interchange within the XML Framework. 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 various tools that can be used for corpus management, as well as with transducers from various XML (and in fact also non-XML) formats into their native formats. This is then another area in which the existing tools can be re-used in the creation and management of the IPI PAN corpus.

3.2. XML Framework: new approach to standardization of text resources

The introduction and widespread of XML and related standards caused a shift of perspective in viewing the issue of standardization and interchange, as applied to linguistic corpora. In 1993, Nancy Ide & Jean Véronis wrote in their proposal for the CES:

“A more powerful way to standardize texts is not only to specify tag syntax, naming rules, etc., but also to identify a relevant set of text categories and specify the actual tags that are to mark these categories in the text together with rules for using these tags. (...) In SGML terms, standardization at this level means that documents of the same type have common DTDs. The largest part of the TEI Guidelines constitutes a standard at this level.”

Ide & Véronis (1993: sect. 4.2)

Notice that the TEI, upon which the CES was based, is extraordinarily successful in this respect: by virtue of its extreme flexibility (rich tagset, modular design, extensibility) it is able to cover an enormous range of the existing schemes used in corpus encoding. At present, however, ensuring a compatible set of markup tags has become much less critical thanks to the powerful transduction methods offered by the XML framework. What is now considered essential for the

⁹ See <http://www.ukoln.ac.uk/metadata/resources/dc/dc-xml-guidelines/> for more concrete examples.

purpose of standardization and interchange is not a common tagset but rather a common **abstract data model**.¹⁰

SGML was a result of a radical change in thinking about document structure: the shift of stress from surface properties of texts to their logical structure and composition. The introduction of XML and the XML Framework appears to have made it possible for that move to proceed a step further: the researchers can now focus on the pure logical design of documents, and the particular tagsets which express this design have a secondary role, and can be abstracted away from and transduced into a common format able to express generalizations about the abstract data model assumed in the given case. This is also a move from semantics provided by multi-page commentaries (cf. the TEI P3/4 Guidelines) to that provided by the machine-readable RDF model, far less ambiguous and clearer than the former, as it is not burdened with the natural language fuzziness and ambiguity.¹¹

In this way, one can talk about a kind of paradigm shift within the approach to the architecture of textual data. The new version of the XCES – not yet distinguished by version number from its late 90's predecessor, but clearly different in its scope and aims – reflects this change by shifting its focus accordingly. As Ide & Romary (2001) claim, the XCES can now be extended to achieve the following:

- support a broad range of annotation types for language data;
- allow multiple annotation levels, where the various annotation levels can be related to each other;
- be open with respect to the information levels and categories within each level;
- allow co-existence of a multitude of coding schemes and standards;
- allow multi-linguality and multi-modality;
- integrate standardization efforts in the US, Europe and Japan;
- provide "off the shelf" and/or easily modifiable XML support for a broad range of annotation types.

This is all made possible by the introduction of the above-mentioned GMT model – a multi-layer model similar to that of the ATLAS architecture but claimed to be even more general (Ide & Romary 2000). This is because unlike ATLAS it does not assume a single internal interchange format, but rather derives its internal format (AML – Annotation Markup Language) for each concrete case, as a result of the interaction of several variables with the underlying meta model of data representation. Once this architecture is released to the public, it will be adopted for the IPI PAN corpus.

4. Structure and preparation of the IPI PAN corpus

This section describes the major tasks that need to be performed in the creation of the IPI PAN corpus. Corpus creation will begin with up-translation of copies of the available texts and creation of the basic metadata (header) information, in as automated way as possible.¹² Next, the texts will undergo a semi-automatic annotation process designed to provide the gross structural markup. After that, the originals will be made read-only. Once this step has been taken, they will be

¹⁰ As Ide & Brew (2000) put it, “a data model is a formalized description of the data objects (in terms of composition, attributes, class membership, applicable procedures, etc.) and relations among them, independent of their instantiation in any particular form.”

¹¹ This is not to diminish the role of the TEI in contrast with the XCES. Recall from footnote 5 that the TEI moves in a very similar direction, for obvious reasons. Thus, the comparison here is to be understood as one between a DTD-based (or SGML-legacy) model and the purely XML-based data model. Besides, the TEI Guidelines are more than merely a description of tag/attribute semantics – they provide an clear introduction to corpus encoding and a user's manual.

¹² Needless to say, the original unmodified versions must be kept at least for cross-checking purposes, or for the purpose of resolving the <hi> tags.

remotely annotated for sentence units and for word-sized tokens that will be further processed by the morphological analyzer and the disambiguator.

4.1. Up-translation

In the initial process of up-translation, all foreign markup has to be removed or replaced with appropriate elements or XML entities. Thus, all rendering information not directly expressed in the unformatted text, such as italics or bold print, has to be expressed by means of the special <hi> tag, which will later be resolved to more specific tags. For the 1st level of CES conformance, it is enough to use the <hi> tag with an appropriate value of the rend attribute. The CES guidelines suggest (in a shortened form) the following values: “bold face”, “boxed”, “italic font”, “roman font”, “underlined”, “capital letters”. More than one value may be specified, if needed. The rend attribute is obligatory on the <hi> tag. It is optional on all other elements (being a member of the a.text attribute class), and should be used there if the borders of the highlighted span of text coincide with the borders of the given element.¹³

A single span of highlighted text may not be marked in a mixed way: if e.g., the highlighted text spans more than a single <p> element, but its beginning (or end) does not coincide with the beginning (or end) of some other <p> element, the <hi> element has to be used throughout, as shown below.

¹³ In fact, as defined in the xcesDoc DTD, the rend attribute is optional everywhere, hence care should be taken to ensure that it is always present on the <hi> element. A minimal change of the DTD would in fact avoid that problem, either by defining the <hi> element separately (although in such cases this single element will be yanked outside of the CES class system, which may turn out to be troublesome if anything in the a.text class is changed in the future), or by defining another attribute class, such as, e.g. a.basetext:

```
<!ENTITY % a.basetext '%a.global;
           wsd          CDATA          #IMPLIED' >
```

then redefining a.text from

```
<!ENTITY % a.text '%a.global;
           rend        CDATA          #IMPLIED
           wsd         CDATA          #IMPLIED' >
```

to

```
<!ENTITY % a.text '%a.basetext;
           rend        CDATA          #IMPLIED' >
```

and changing the attribute structure of the <hi> element from

```
<!ATTLIST hi %a.text; >
```

to

```
<!ATTLIST hi %a.basetext;
           rend        CDATA          #REQUIRED>
```

A separate question is whether the addition of another class (thus changing the flat class structure of the CES) is warranted in this case. A schema implementation would avoid this issue, by simply requiring a change to the single element <hi> instead of affecting the entire document structure definition.

(2) `<p>...<hi rend="it">...</hi></p><p><hi rend="it">...</hi></p>` (correct)

`<p>...<hi rend="it">...</hi></p><p rend="it">...</p>` (incorrect)

This is again a matter of guaranteeing the proper retrieval of the relevant information.

All character encoding in the text constituting the corpus will also have to be unified. In order to facilitate readability, the target character set will be Latin-2 (ISO-8859-2), as XML engines perform on-the-fly translation from Latin-2 to Unicode (the native XML character encoding), and because Unicode is not yet widespread in all of the commonly used computer systems.¹⁴ In most cases, foreign markup removal/replacement as well as character set conversion can be performed by means of Perl or shell scripts, many of which have been made publicly available by e.g. the MULTEXT, MULTEXT-EAST or LT-XML projects. All XML processors are required to translate the sequence of carriage-return (CR) and the line-feed (LF) characters, as well as any CR not followed by LF into a single LF character (see <http://www.w3.org/TR/REC-xml#sec-line-ends>). In this way, Windows line breaks (LF-CR) as well as MacOS line breaks (CR) are uniformly translated into the Unix format (LF). Another automated task is element occurrence count, needed for the header. This can also be easily achieved thanks to a Perl script freely available from the CES site.

At the end of the initial process of up-translation, the original character sequence comprising the document must be retained. In particular, none of the original sequence of characters may be altered.¹⁵ The original data should only appear as element content, and never in attributes (the TEI allows that in some cases, see Ide and Véronis 1993). No other data may appear as element content. The original order of the data should not be changed.¹⁶ Rendition information may be exempt from the above conditions if its sole purpose would be to recreate the original printed source. This concerns e.g. non-meaningful page breaks.

Later, all kinds of quotes will need to be converted to the appropriate tags. Here, attention should be drawn to the distinction the CES makes between the `<quote>` and `<q>` elements. The former is used for the so-called long quotes, whereas the latter for quotes contained within the paragraph level.

All the up-translation procedures have to be described within the `<encodingDesc>` section of the text and/or corpus header.

4.2. The original document

The following are the two possible expansions of the `<cesCorpus>` element:

¹⁴ This concerns translation from e.g. the Windows code page; wherever it is essential that the text in the corpus remains in a non-Latin-2 encoding, it has to be identified by the `wsc` attribute within the corpus and by the `<writingSystem>` element in the header.

¹⁵ This concerns the main text as opposed to footnotes, titles, captions, bibliographic references, etc. which are considered auxiliary text and may often be exempt from linguistic analysis.

¹⁶ This poses an interesting problem with regard to where footnotes fit into the data stream, if they are not ignored as auxiliary text – whether they should be included at the point of reference or at the end of the containing section or chapter, or the entire text.

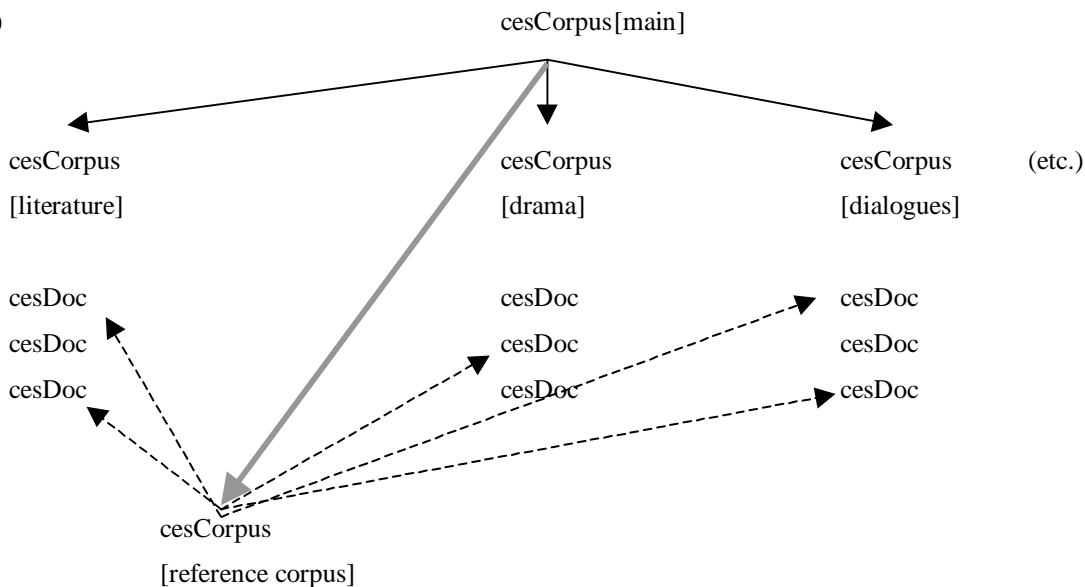
(3)

```
a. <cesCorpus>
    <cesHeader version="">
    </cesHeader>
    <cesDoc version="">           [one or more]
    </cesDoc>
</cesCorpus>
```

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

At this point, we envision only one-level deep nesting of the `<cesCorpus>` elements, according to the primary divisions according to the genre of the text included there. 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.

(4)



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, this subcorpus belongs on a different plane, which should be clearly indicated in the appropriate headers.

The near-minimal structure required by the `xcesDoc` DTD is as follows.

(5)

```
<cesDoc version="3.19">
  <cesHeader version="2.1"> ... </cesHeader>
  <text>
    <body>
      <bibl/>17 [optional, repeatable]
      <div type=""> [optional, repeatable, self-nesting]
        <opener/> [optional, repeatable]
        <sp> [optional, repeatable, for spoken texts]
          <p/> [optional, repeatable]
        </sp>
        <closer/> [optional, repeatable]
      </div>
    </body>
  </text>
</cesDoc>
```

The `version` attribute on the `<cesDoc>` element is the version of the compliant DTD. On the `<cesHeader>` element, it is the version number of the `xheader.elt` file containing the definitions for header contents.

4.3. The extra-textual information annotation

While the kind of markup referred to in the preceding section identifies certain primarily structural facts about the given document, standard corpora make crucial use of two kinds of extra-textual information: metadata (containing information about the document provenance, its content, the annotation scheme used, the entity that created the annotation information, etc.) and morphosyntactic annotation. We will have little to say about either of these kinds of information as they are the subjects of other reports. In the present report, we restrict ourselves to a brief discussion of what the XCES requires of these two kinds of annotation in order for them to qualify as proper parts of the corpus, skipping any discussion of their actual content.

4.3.1. Metadata: the header

The corpus header as well as the headers for subcorpora and individual texts will be included in the main text by means of XML external entities. This way, the header information will be available for modifications while the base text will remain in read-only form.

In the preceding sections, we have referred to some of the crucial elements of the XCES header. See http://www.mpi.nl/world/ISLE/overview/Overview_CES.html for a graphical overview of the header structure of the CES; a verbose description is provided in chapter 3 of the CES guidelines, at <http://www.cs.vassar.edu/CES/CES1-3.html>.

4.3.2. Morphosyntactic annotation: remote markup DTDs

The stand-off annotation files instantiate the `xcesAna` DTD. They provide elements for gross divisions within the main

¹⁷ This kind of element marking is used here for the sake of brevity and it does not imply that the element in question is empty.

document, with `<chunkList>` corresponding roughly to the main text of the document (or selected parts thereof), `<chunk>` to e.g. chapter or section divisions (marked by the appropriate values of the `type` attribute), and `<par>` corresponding to paragraph content. The `<par>` element contains `<s>` (for sentence segmentation) and `<tok>` (for word-sized chunks of text). The fine structure of such files is described at <http://www.cs.vassar.edu/CES/CES1-5.html>. The basic structure of an annotation file is presented below.

(6)

```

<cesAna version="1.11">
  <cesHeader version="2.1">      [optional]
  </cesHeader>

  <chunkList>
    <chunk>                        [repeatable]
      <par>                          [optional, repeatable]
        <s>                            [optional, repeatable]
          <tok>                          [repeatable]
            <orth/>

            <disamb>                    [repeatable (sic!), optional]
              <ctag/>
              <msd/>
            </disamb>

            <lex>                        [repeatable, optional]
              <base/>
              <msd/>
              <ctag/>
            </lex>

          </tok>
        </s>
      </par>
    </chunk>
  </chunkList>

</cesAna>

```

For the IPI PAN corpus, two kinds of stand-off annotation documents are initially planned. The first kind will contain sentence segmentation for the given text. It will refer to the base text via one-way links. The other kind of remote markup documents will contain morphosyntactic annotation. This kind of annotation will not reference the base text but rather the document with `<s>`-alignment annotation, as illustrated below. The sentence alignment document will have the `version` attribute of the `<cesAna>` element set to "sent", whereas the POS alignment document will be 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.

(7) original document ← sentence alignment ← POS annotation

In this way, the sentence alignment documents in the middle act as the base for various possible POS annotation

documents, as well as documents containing e.g. syntactic annotation, which may be added at a later time.¹⁸ Notice for example that one POS annotation document may be marked as “tok lex” and another as “tok disamb”, should such division be necessary, for example for the purpose of comparing the efficiency and precision of various taggers.

This kind of architecture 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 alignment documents (instantiating the third kind of XCES DTD, namely the xcesAlign DTD) will contain two-way links addressing the appropriate sentence alignment documents. 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 the two 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.¹⁹ Turning to matters more sublime, we may also mention the numerous translations of e.g. Shakespeare’s plays, among many others.

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 would 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*.²⁰ By locating tokenization 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.

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.

¹⁸ As discussed by Woliński & Przepiórkowski (2001), disambiguation in the basic version of the IPI PAN corpus morphosyntactic annotation will be restricted to the domain of the <s>.

¹⁹ 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.

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 elsewhere. The more so that this raises potential problems when it comes to e.g. correcting typographical errors in the original 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.

²⁰ I am grateful to Adam Przepiórkowski for a discussion on the issue of radical tokenization.

References

- Bański, Piotr (2000). Morphological and phonological analysis of auxiliary clitics in Polish and English. PhD. Dissertation, University of Warsaw.
- Bird, Steven, and Mark Liberman (2001). "A formal framework for linguistic annotation. *Speech Communication*, 33, 23-60.
- Dębowski, Łukasz (2001). Tagowanie i dezambiguacja syntaktyczna. Przegląd metod i oprogramowania. Technical report, Polish Academy of Sciences.
- Ide, Nancy, Priest-Dorman, Greg, and Jean Véronis (1996). Corpus Encoding Standard. Available at <<http://www.cs.vassar.edu/CES/>>.
- Ide, Nancy and Chris Brew (2000). Requirements, Tools, and Architectures for Annotated Corpora. *Proceedings of Data Architectures and Software Support for Large Corpora*. Paris: European Language Resources Association, 1-5. Available from <<http://www.cs.vassar.edu/faculty/ide/pubs.html>>.
- 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. Available from <<http://www.cs.vassar.edu/faculty/ide/pubs.html>>.
- Ide, Nancy, and Laurent Romary (2000). "Encoding syntactic annotation". Chapter 10 of *Treebanks: Building and using syntactically annotated corpora*, ed. Anne Abeillé. Amsterdam: Kluwer Academic Publishers. Also available from <<http://treebank.linguist.jussieu.fr/toc.html>>.
- Ide, Nancy, and Laurent Romary (2001). Standards for Language Resources. Department of Computer Science, Vassar College and Equipe Langue et Dialogue, LORIA/INRIA.
- Kurcz, I., Lewicki, A., Sambor, J., Szafran, K., and J. Woronczak (1990). Słownik frekwencyjny polszczyzny współczesnej. Kraków: Instytut Języka Polskiego PAN.
- Kupść, Anna and Elżbieta Hajnicz (2001). Przegląd analizatorów morfologicznych dla języka polskiego. Technical report, Polish Academy of Sciences.
- Woliński, Marcin and Adam Przepiórkowski (2001). Projekt sposobu morfosyntaktycznego anotowania korpusu języka polskiego. Technical report, Polish Academy of Sciences.

Annex 1: Important URLs

ANC (American National Corpus): <http://www.cs.vassar.edu/~ide/anc>

ATLAS (Architecture and Tools for Linguistic Analysis Systems):

<http://www.nist.gov/speech/atlas/index.html>

BNC (British National Corpus): <http://info.ox.ac.uk/bnc/>

CES (Corpus Encoding Standard): <http://www.cs.vassar.edu/CES/>

EAGLES (Expert Advisory Group on Language Engineering Standards):

<http://www.ilc.pi.cnr.it/EAGLES/home.html>

ISLE Metadata Initiative: <http://www.mpi.nl/world/ISLE/>

ISLE (International Standards for Language Engineering):

http://lingue.ilc.pi.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm

LT XML: <http://www.ltg.ed.ac.uk/software/xml/>

MATE (Multilevel Annotation, Tools Engineering): <http://mate.nis.sdu.dk/>

MULTEXT (Multilingual Text Tools and Corpora):

<http://www.lpl.univ-aix.fr/projects/multext/>

MULTEXT-EAST (Multi-lingual Texts and Tools for Central and Eastern European

Languages): <http://nl.ijs.si/ME/>

RELAX NG: <http://www.oasis-open.org/committees/relax-ng/>

Schematron: <http://www.ascc.net/xml/resource/schematron/schematron.html>

TEI (Text Encoding Initiative): <http://www.tei-c.org/>

List of projects based on the TEI: <http://www.tei-c.org/Applications/index.html>

TEI-Lite: <http://www.tei-c.org/Lite/index.html>

TIPSTER annotation repository: <http://crl.nmsu.edu/Research/Projects/tipster/annotation/>

XML Schema (<http://www.w3.org/TR/xmlschema-0>, <http://www.w3.org/TR/xmlschema-1>, <http://www.w3.org/TR/xmlschema-2>)

Annex 2: the xcesDoc DTD

```
<!-- -->
<!-- -->
<!-- Corpus Encoding Standard -->
<!-- -->
<!-- CES -->
<!-- -->
<!-- Encoding conventions for level 1 -->
<!-- -->
<!-- -->
<!-- -->
<!-- -->
<!-- -->
<!-- -->
<!-- -->
    $Date: 1996/06/13 13:16:45 $
    $Revision: 3.19 $
-->
<!-- -->
<!-- ENTITY DECLARATIONS -->
<!-- -->
<!-- Global attributes -->
<!ENTITY % a.global '
    id          ID          #IMPLIED
    n           CDATA       #IMPLIED
    xml:lang    CDATA       #IMPLIED
    lang        IDREF       #IMPLIED' >
<!ENTITY % a.text '%a.global;
    rend        CDATA       #IMPLIED
    wsd         CDATA       #IMPLIED' >
<!-- Elements that can appear between paragraphs -->
<!ENTITY % m.inter ' bibl | quote | list |
    poem | note | caption | figure | table ' >
<!-- Sub-paragraph elements -->
<!ENTITY % m.token 'abbr | date | num |measure |
    name | term | time |' >
<!ENTITY % m.phrase '%m.token; foreign | mentioned |
    distinct | title | hi | list |
    corr | gap | reg | ptr | ref' >
<!ENTITY % m.phrase.hi '%m.token; foreign | mentioned |
    distinct | title | list |
    corr | gap | reg | ptr | ref' >
<!ENTITY % m.phrase.foreign '%m.token; mentioned |
    distinct | title | hi | list |
    corr | gap | reg | ptr | ref' >
<!ENTITY % m.phrase.distinct '%m.token; foreign | mentioned |
    title | hi | list |
    corr | gap | reg | ptr | ref' >
<!ENTITY % m.phrase.mentioned '%m.token; foreign |
    distinct | title | hi | list |
```

```

                corr | gap | reg | ptr | ref'                >
<!ENTITY % m.phrase.title '%m.token; foreign | mentioned |
                distinct | hi | list |
                corr | gap | reg | ptr | ref'                >
<!--                Content model declarations                -->
<!ENTITY % base.seq    '#PCDATA | num | abbr'              >
<!ENTITY % phrase.seq  '#PCDATA | %m.phrase;'              >
<!ENTITY % phrase.hi.seq '#PCDATA | %m.phrase.hi;'         >
<!ENTITY % phrase.foreign.seq '#PCDATA | %m.phrase.foreign;' >
<!ENTITY % phrase.distinct.seq '#PCDATA | %m.phrase.distinct;' >
<!ENTITY % phrase.mentioned.seq '#PCDATA | %m.phrase.mentioned;' >
<!ENTITY % phrase.title.seq '#PCDATA | %m.phrase.title;'   >
<!ENTITY % par.seq     '(%m.inter; | p | sp )*'            >
<!ENTITY % subpar.seq  '%phrase.seq; | s | q'              >

<!--                -->
<!--                ELEMENT DECLARATIONS                -->
<!--                -->
<!--                HIGH-LEVEL COMPONENTS                -->

<!ELEMENT cesCorpus    (cesHeader,(cesDoc+ | cesCorpus+)) >
<!ATTLIST cesCorpus
    %a.global;
    TEIform          CDATA          'teiCorpus.2' >

<!ELEMENT cesDoc       (cesHeader, text)                  >
<!ATTLIST cesDoc
    %a.global;
    type             CDATA          "text"
    version          CDATA          #REQUIRED
    TEIform          CDATA          'TEI.2' >

<!ENTITY % xces.header SYSTEM 'xheader.elt'                >
%xces.header;

<!--                WRITTEN TEXTS                -->

<!ELEMENT text         (body | group)                      >
<!ATTLIST text
    %a.global;
    complete          (y | n )          "y"
    decls             IDREFS           #IMPLIED >

<!ELEMENT body         (%par.seq; , div*)                  >
<!ATTLIST body
    %a.text;
    decls             IDREFS           #IMPLIED >

```

```

<!ELEMENT group          (%par.seq; , body+)          >
<!ATTLIST group
  decls                    IDREFS                    #IMPLIED      >

<!ELEMENT div            ((opener | head | byline)*,
  (((p | sp | %m.inter;)+, div*)
  | div+), (closer | byline)* )          >
<!ATTLIST div
  complete                 (y | n)                  "y"
  type                     CDATA                     #REQUIRED
  decls                    IDREFS                    #IMPLIED      >

<!--                    Opening elements                -->

<!ELEMENT opener        (%phrase.seq; | dateline | keywords)* >
<!ATTLIST opener
  %a.text;                  >

<!ELEMENT head          (%phrase.seq;)*                >
<!ATTLIST head
  %a.text;
  type                     ( byline | display |
  attached | unspec )      "unspec"          >

<!--                    Keyword lists, bylines, datelines  -->

<!ELEMENT keywords      (term+ | list)                >
<!ATTLIST keywords
  %a.text;
  scheme                    IDREF                    #IMPLIED      >

<!ELEMENT byline        (%phrase.seq; | docAuthor)*    >
<!ATTLIST byline
  %a.text;                  >

<!ELEMENT docAuthor     (%base.seq;)*                >
<!ATTLIST docAuthor
  %a.text;                  >

<!ELEMENT dateline      (%base.seq; | date | time | name | address)* >
<!ATTLIST dateline
  %a.text;                  >

<!ELEMENT address       (%base.seq;)*                >
<!ATTLIST address
  %a.text;                  >

<!--                    Closing element                  -->

<!ELEMENT closer        (%phrase.seq; | dateline | keywords)* >
<!ATTLIST closer
  %a.text;                  >

<!--                    PARAGRAPH-LEVEL ELEMENTS  THE CLASS M.INTER  -->

<!--                    Written paragraphs                -->

<!ELEMENT p             (%subpar.seq;)*                >
<!ATTLIST p
  %a.text;                  >

<!--                    Quotations                      -->

```

```

<!ELEMENT quote      (%subpar.seq; | p)*          >
<!ATTLIST quote
  type                CDATA                    #IMPLIED    >

<!--                Lists                -->

<!ELEMENT list      (head?, (item+ | (label, item)+)) >
<!ATTLIST list
  %a.text;            >

<!ELEMENT item      (%subpar.seq; | p)*          >
<!ATTLIST item
  %a.text;            >

<!ELEMENT label     (%phrase.seq;)*              >
<!ATTLIST label
  %a.text;            >

<!--                Annotations          -->

<!ELEMENT note      (%subpar.seq; | p)*          >
<!ATTLIST note
  place              (side | foot | end | unspec)
                    "unspec"                  >

<!ELEMENT bibl      (%phrase.seq; | author)*     >
<!ATTLIST bibl
  %a.text;            >
  %a.declarable;     >

<!ELEMENT author    (%base.seq;)*               >
<!ATTLIST author
  %a.text;            >

<!--                Poems                -->

<!ELEMENT poem      (head?, (lg | l )+ )         >
<!ATTLIST poem
  %a.text;            >

<!ELEMENT lg        (l | lg)+                   >
<!ATTLIST lg
  %a.text;            >
  type                CDATA                    #IMPLIED    >
  part               (y | n | u)                "u"         >

<!ELEMENT l         (%phrase.seq;)*              >
<!ATTLIST l
  %a.text;            >
  part               (y | n | u)                "u"         >

<!--                Figures              -->

<!ELEMENT figure    (head?, p*, figDesc?, text?) >
<!ATTLIST figure
  figure            %a.text;                    >
  entity            ENTITY                    #IMPLIED    >

<!ELEMENT figDesc   (%phrase.seq;)*              >
<!ATTLIST figDesc
  %a.text;            >

<!--                Tables                -->

<!ELEMENT table     (head?, row+)                >

```

```

<!ATTLIST table
    rows          NMTOKEN          #IMPLIED
    cols          NMTOKEN          #IMPLIED
>

<!ELEMENT row    (cell | table)+
<!ATTLIST row
    role          CDATA            "data"
>

<!ELEMENT cell   (%phrase.seq;)*
<!ATTLIST cell
    role          CDATA            "data"
    rows          NMTOKEN          "1"
    cols          NMTOKEN          "1"
>

<!--           Captions           -->

<!ELEMENT caption (%phrase.seq;)*
<!ATTLIST caption
    %a.text;
>

<!--           Transcriptions of dialogues, speeches, debates,
<!--           interviews, etc., and drama           -->

<!ELEMENT sp     (speaker | p | stage)+>
<!ATTLIST sp
    who           NMTOKEN          #IMPLIED
>

<!ELEMENT speaker (%base.seq;)*
<!ATTLIST speaker
    %a.text;
>

<!ELEMENT stage  (%base.seq;)*
<!ATTLIST stage
    %a.text;
    type         CDATA            #IMPLIED
>

<!--           SENTENCES, QUOTED DIALOGUE WITHIN PARAGRAPHS           -->

<!ELEMENT s      (%subpar.seq;)*
<!ATTLIST s
    next         IDREF            #IMPLIED
    prev         IDREF            #IMPLIED
    type         CDATA            #IMPLIED
    broken       (yes | no)       "no"
>

<!ELEMENT q      (%subpar.seq;)*
<!ATTLIST q
    next         IDREF            #IMPLIED
    prev         IDREF            #IMPLIED
    type         CDATA            #IMPLIED
    direct       (y | n | unspecified)
                "unspecified"
    who         CDATA            #IMPLIED
    broken       (yes | no)       "no"
>

<!--           PHRASE-LEVEL ELEMENTS   THE CLASS M.PHRASE           -->

```

```

<!--          Editorial Changes          -->

<!ELEMENT gap          EMPTY          >
<!ATTLIST gap          %a.text;
          desc          CDATA          #IMPLIED
          reason       CDATA          #IMPLIED
          resp         CDATA          #IMPLIED
          cert         CDATA          #IMPLIED          >

<!ELEMENT reg          (%phrase.seq;)*          >
<!ATTLIST reg          %a.text;
          orig         CDATA          #IMPLIED
          resp         CDATA          #IMPLIED
          cert         CDATA          #IMPLIED          >

<!ELEMENT corr        (%phrase.seq;)*          >
<!ATTLIST corr        %a.text;
          sic          CDATA          #IMPLIED
          resp         CDATA          #IMPLIED
          cert         CDATA          #IMPLIED          >

<!--          Highlighted text          -->

<!ELEMENT hi          (%phrase.hi.seq;)*>
<!ATTLIST hi          %a.text;          >

<!--          Other Phrase-level Elements          -->

<!ELEMENT date        (%base.seq;)*          >
<!ATTLIST date        %a.text;
          ISO8601      CDATA          #IMPLIED          >

<!ELEMENT foreign     (%phrase.foreign.seq;)*          >
<!ATTLIST foreign     %a.text;          >

<!ELEMENT distinct    (%phrase.distinct.seq;)*          >
<!ATTLIST distinct    %a.text;
          type         CDATA          #IMPLIED          >

<!ELEMENT mentioned   (%phrase.mentioned.seq;)*          >
<!ATTLIST mentioned   %a.text;          >

<!ELEMENT measure     (%base.seq;)*          >
<!ATTLIST measure     %a.text;
          type         (weight | length | count | area | volume | currency)
#IMPLIED
          value        CDATA          #IMPLIED          >

<!ELEMENT name        (%base.seq;)*          >
<!ATTLIST name        %a.text;
          type         ( person | place | org )
#IMPLIED          >

<!ELEMENT term        (%base.seq;)*          >
<!ATTLIST term        %a.text;
          type         CDATA          #IMPLIED          >

```

```

<!ELEMENT time          (%base.seq;)*          >
<!ATTLIST time
  ISO8601          CDATA          #IMPLIED
  type            (am | pm | 24hour | descriptive) #IMPLIED
>

<!ELEMENT title        (%phrase.title.seq;)*   >
<!ATTLIST title
  type            CDATA          #IMPLIED
>

<!ELEMENT abbr         (#PCDATA)              >
<!ATTLIST abbr
  expan          CDATA          #IMPLIED
  resp           IDREF         #IMPLIED
  cert           CDATA          #IMPLIED
  type           CDATA          #IMPLIED
>

<!ELEMENT num         (#PCDATA)              >
<!ATTLIST num
  type           CDATA          #IMPLIED
  value          CDATA          #IMPLIED
>

<!--                SEGMENTATION, LINKING, ALIGNMENT          -->

<!--                Simple cross references                    -->

<!ELEMENT ptr         EMPTY                  >
<!ATTLIST ptr
  corresp        IDREFS         #IMPLIED
  next           IDREF         #IMPLIED
  prev           IDREF         #IMPLIED
  type           CDATA          #IMPLIED
  resp           CDATA          #IMPLIED
  crdate         CDATA          #IMPLIED
  targType       NMTOKENS      #IMPLIED
  targOrder     (y | n | u)    "u"
  evaluate       (all | one | none) #IMPLIED
  target         IDREFS         #REQUIRED
>

<!ELEMENT ref         (%phrase.seq;)*       >
<!ATTLIST ref
  corresp        IDREFS         #IMPLIED
  next           IDREF         #IMPLIED
  prev           IDREF         #IMPLIED
  type           CDATA          #IMPLIED
  resp           CDATA          #IMPLIED
  crdate         CDATA          #IMPLIED
  targType       NMTOKENS      #IMPLIED
  targOrder     (y | n | u)    "u"
  evaluate       (all | one | none) #IMPLIED
  target         IDREFS         #IMPLIED
>

```

Annex 3: the xcesAna DTD

```

<!-- -->
<!-- -->
<!-- Corpus Encoding Standard -->
<!-- -->
<!-- CES -->
<!-- -->
<!-- Encoding conventions for annotated data -->
<!-- -->
<!-- -->
<!-- -->
<!-- -->
<!-- -->
    $Date: 1996/08/05 19:07:30 $
    $Revision: 1.11 $
-->
<!-- -->
<!-- Global attributes -->

<!ENTITY % a.global '
    id ID #IMPLIED
    n CDATA #IMPLIED
    xml:lang CDATA #IMPLIED
    lang IDREF #IMPLIED' >

<!ENTITY % a.ana '%a.global;
    type CDATA #IMPLIED
    ws CDATA #IMPLIED' >

<!ELEMENT cesAna (cesHeader?, chunkList) >
<!ATTLIST cesAna %a.ana;
    doc CDATA #IMPLIED
    version CDATA #REQUIRED >

<!ENTITY % xces.header SYSTEM 'xheader.elt' >
%xces.header;

<!ELEMENT chunkList (chunk)+ >
<!ATTLIST chunkList %a.ana; >

<!ELEMENT chunk (tok+ | s+ | par+) >
<!ATTLIST chunk %a.ana;
    doc CDATA #IMPLIED
    from CDATA #IMPLIED
    to CDATA #IMPLIED >

<!ELEMENT par (tok | s)+ >
<!ATTLIST par %a.ana;
    from CDATA #IMPLIED
    to CDATA #IMPLIED >

<!ELEMENT s (#PCDATA | tok | s)* >
<!ATTLIST s %a.ana;
    from CDATA #IMPLIED
    to CDATA #IMPLIED
    next IDREF #IMPLIED
    prev IDREF #IMPLIED
    broken (yes | no) "no" >

```

```

<!ELEMENT tok          (orth, ((disamb*, lex*)
                             | (base?, ctag?, msd?)))      >
<!ATTLIST tok
  class                CDATA          #IMPLIED
  from                 CDATA          #IMPLIED
  to                   CDATA          #IMPLIED      >

<!ELEMENT orth        (#PCDATA)          >
<!ATTLIST orth
  %a.ana;              > >

<!ELEMENT disamb      (ctag, msd?)+      >
<!ATTLIST disamb
  %a.ana;              > >

<!ELEMENT lex         (base, msd?, ctag)  >
<!ATTLIST lex
  %a.ana;              > >

<!ELEMENT base        (#PCDATA)          >
<!ATTLIST base
  %a.ana;              > >

<!ELEMENT msd         (#PCDATA)          >
<!ATTLIST msd
  %a.ana;              > >

<!ELEMENT ctag        (#PCDATA)          >
<!ATTLIST ctag
  %a.ana;              >
  certainty            CDATA          #IMPLIED      >

```

Pracę zgłosił: Piotr Bański

Adres autora

Instytut Anglistyki,
Uniwersytet Warszawski
Nowy Świat 4
00-497 Warszawa

Symbol klasyfikacji rzeczowej

Na prawach rękopisu
Printed as a manuscript