The TEI and the NCP: the model and its application

Piotr Bański
Institute of English Studies
University of Warsaw
pkbanski@uw.edu.pl

Adam Przepiórkowski
Institute of Computer Science
Polish Academy of Sciences
adamp@ipipan.waw.pl

Abstract

We present the National Corpus of Polish (NCP), a TEI-encoded 1-billion-word text corpus with multiple layers of linguistic annotation, the product of co-operation of a consortium of all the major Polish institutions that created their own significant corpora in the past. We review the major properties of the corpus and of its architecture, with an eye to the hot topics of today: interoperability and sustainability. Special attention is paid to the status of the encoding schemes of the corpus vis-à-vis the currently popular annotation standards.

1. Introduction: all the buzzwords

The interrelated issues of sustainability and interoperability are part of the landscape of any system of representation and processing of linguistic knowledge. Helbig (2001), quoted in (Witt *et al.*, 2009:7), lists interoperability, homogeneity, and communicability as three requirements for such a system. Helbig's *homogeneity* requires the same formalism across different levels of linguistic description, and *communicability* refers to the documentation, conditioning successful teamwork and allowing to share resources among different teams.

Our primary focus here is on a particular subset of knowledge inherent in *Language Resources* (LRs)[1], and more specifically, in what Witt *et al.*, (2009) call *static text-based LRs*, i.e., language corpora, although we will also mention *dynamic LRs*, that is tools that manipulate or query corpora. In this context, *interoperability* means, generally, the ability of LRs to "understand" each other and to interact. As Ide (2010) points out, this can take place at several levels, notably at the syntactic level, e.g. via an abstract pivot format that makes it possible to reduce the number of mappings between schemas (cf. also (Ide and Romary, 2007)), and at the semantic level, by reference to a common data model and, crucially, a shared inventory of reference categories.

It is a trivial observation that the practical value of LRs lies in their use, possibly recurrent, and ideally permanent. This takes us towards the concept of *sustainability*, i.e., (very generally and imprecisely) the ability to ensure a prolonged (and ideally permanent) use of LRs. Sustainability, as defined by e.g. (Nathan, 2006) or (Simons and Bird, 2008) requires that a LR be (i) extant (Nathan: permanent), (ii) usable (Nathan: proficiently prepared), and (iii) relevant (Nathan: pertinent). Simons and Bird (2008), whose terminology we adopt here, additionally split usability into sub-conditions: discoverability, availability, interpretability, and portability. Note that the last two properties provide for interoperability.

We will not provide a thorough overview of all these requirements, merely noting them as we proceed in our presentation of the largest linguistically annotated text corpus of Modern Polish, the National Corpus of Polish (NCP, also known under its native abbreviation, NKJP, http://nkjp.pl/).[2] In doing so, we briefly report on the origin and the nearest future of the corpus as well as the design decisions that shaped it.

2. National Corpus of Polish: history and future

The NCP is the deliverable of project number R17 003 03, sponsored by the Polish Ministry of Science and Higher Education. It was launched in December 2007 and will terminate at the end of 2010. The project is carried out by a consortium of four institutions that developed their own significant corpora in the past. These corpora are now joined into a single resource that has been expanded to nearly three times the size of the original corpora put together. The project members are the following:

- the Institute of Computer Science at the Polish Academy of Sciences in Warsaw (ICS PAS),
- the Institute of the Polish Language at the Polish Academy of Sciences in Cracow (IPL PAS),
- the PWN Scientific Publishers in Warsaw,
- the PELCRA group at the University of Łódź.

These four institutions have combined their expertise to merge, uniformly encode, enlarge, and enhance their resources, eventually producing a 1-billion-word corpus (with a carefully balanced 300-million-word subpart), annotated for various levels of linguistic description and designed to last and serve current and future research (for more details, see Przepiórkowski *et al.*, 2010).

After the project has completed at the end of 2010, the

[1] A Language Resource is "any physical or digital item that is a product of language documentation, description, or development, or is a tool that specifically supports the creation and use of such products" (Simons and Bird, 2008). See (Witt *et al.*, 2009) for a suggested taxonomy of LRs.

[2] "NKJP" expands into "Narodowy Korpus Języka Polskiego".

NCP will be used as the empirical basis for the development of a new large dictionary of Modern Polish that is being created at IPL PAS. 1-million-word demo of the corpus may be released under an open license, pending the solving of licensing issues.

The contents of this section, apart from setting the context for the rest of the paper, address some of the requirements mentioned in section 1, for example the requirement of *relevance*: the NCP is the first corpus of its kind in Poland, and the first such corpus of Polish. To our knowledge, it is also the first corpus of this size (10^9 words) with homogeneous encoding of multiple hierarchical layers of linguistic annotation (to be reviewed below), in the world. The entire bulk of the corpus (though, understandably, not the entire set of annotations, some of which are still being created) is already available for searching via two interfaces. The NCP will have a large, carefully balanced subcorpus and it contains nearly 2 million words if informal conversational Polish, which is precious for two reasons: firstly, there has been no corpus of (transcribed) spoken Polish of such a size before, and secondly, most of the digital data is still available for being aligned with the recordings, which opens a further exciting research perspective.

We also address the issue of *permanence*: copies of the corpus are made regularly and its nearest future is secured. Attention is paid to the question of its long-term persistence, which will be reported on in due time.

As for *availability*, the corpus may not be released in the source form due to the numerous legacy restrictions on the use of the data that it contains: many texts have been released to the NKJP Consortium on the condition that they are not distributed further. However, the corpus may already be queried in its entirety, and a 1-million-word part of it (composed of texts carefully selected for their lack of copyright restrictions) will most probably be released publicly – we wish to note that this is much, much more than what many other closed corpora release. *Discoverability* of the corpus is already partially taken care of (it is, naturally, part of the LREC LR Map), and will also be addressed after the project is completed. We look at sustainability and interoperability of the NCP in sections 3 through 5 below.

3. Architecture: stand-off annotation

The NCP is built according to the guidelines for annotating modern LRs and uses the so-called stand-off mechanism of annotation (Thompson and McKelvie, 1997; Ide and Romary, 2007), whereby each annotation document (typically, though not always, containing information pertaining to a single level of grammatical description) is located in a separate file that references another annotation file or the source text by means of various pointing mechanisms. The typical contents of a leaf directory in the corpus are as presented in the list below (see http://nlp.ipipan.waw.pl/TEI4NKJP/ for working versions of these files; the file NKJP_header.xml belongs here only virtually – we look at its role presently):

- text_structure.xml
- ann_segmentation.xml
- ann_morphosyntax.xml
- ann_senses.xml
- ann_words.xml
- ann_named.xml
- ann_groups.xml
- header.xml
- (NKJP_header.xml)

Above, the file text_structure.xml stores the source text – this file contains coarse-grained inline structural annotation, typically down to the paragraph level. The other files contain annotations of other kinds, organized in a hierarchy: the first is ann_segmentation.xml, containing the segmentation layer that identifies the sentence boundaries and the contiguous non-overlapping sequence of individual segments (including segmental ambiguities, cf. (Bański and Przepiórkowski, 2009)), by addressing character spans in the source text.[3] The segmentation layer can be the pivot layer for many other annotation documents, depending on the setup of the particular corpus and the nature of annotations. In the NCP, however, only the morphosyntactic layer (ann_morphosyntax.xml) is built on top of it. This layer contains all the possible morphosyntactic interpretations of each segment together with an optional disambiguation section that points at the most likely interpretation. The morphosyntactic layer serves as the basis for three other layers, namely those (a) identifying syntactic words (ann_words.xml), (b) identifying named entities (ann_named.xml, cf. (Savary *et al.*, 2010)), and (c) disambiguating selected polysemic lexemes (ann_senses.xml, cf. (Młodzki and Przepiórkowski, 2009)). Finally, the level of syntactic chunks (ann_groups.xml, cf. (Głowińska and Przepiórkowski, 2010)) references the syntactic word level. The file header.xml is the local TEI header, included by all the other files in the directory, whether containing the source text or the annotations. The file NKJP_header.xml is the main corpus header, included by all the files in the entire corpus and thus binding them into a single virtual unit.

It has to be pointed out that stand-off architecture is one of the preconditions for *sustainability* and *interoperability*. A stand-off annotated LR preserves the source text in a minimally marked-up form and hence as capable of being easily extracted or processed by future versions of the current tools or by new tools. Such a resource is also easily expandable, which also adds to its attractiveness (Ide and Romary, 2007). Interoperability of stand-off annotated resources can be realised both at the level of the source text and at the level(s) of the annotation layers: it becomes possible to e.g. compare tagsets, conflicting

[3] These spans can be smaller than orthographic words – for the motivation see (Bański and Przepiórkowski, 2009), for their treatment at the level of syntactic words (ann_words.xml), see (Głowińska and Przepiórkowski, 2010).

annotations, or outputs of different tools; it is much easier to map the contents of annotation layers onto different resources. The criterion of heterogeneity becomes important in this regard (Witt et al., 2009), and we shall see in the next section that the NCP fulfils it.

4. Encoding format: Text Encoding Initiative XML

The NCP is encoded in the popular TEI XML encoding standard (TEI Consortium, 2010), a *de facto* standard for resources of many kinds used in the Humanities and for LRs in general.[4] The TEI Guidelines provide a variety of means to encode linguistic information in LRs. When tailoring the TEI model for the NCP, we attempted to follow the existing standards for linguistic annotation. That task was not difficult because of the origin of many of these standards. The current standards that have been or are being established by ISO TC 37 SC 4 committee (http://www.tc37sc4.org/), known together as the LAF (Linguistic Annotation Framework) family of standards, cf. (Ide and Romary, 2007), descend in part from an early application of the TEI, back when the TEI was still an SGML-based standard. That application was the Corpus Encoding Standard (Ide, 1998), later redone in XML and known as XCES (Ide *et al.*, 2000). XCES was a conceptual predecessor of the current ISO LAF pivot format for syntactic interoperability of annotation formats, GrAF (Graph Annotation Framework, (Ide and Suderman, 2007)). GrAF defines an XML serialization of the LAF data model consisting of directed acyclic graphs with annotations (also expressible as graphs), attached to nodes. This basic data model is in fact common to the TEI formats defined for the NCP, the LAF family of standards, and the other standards and best practices such as Tiger-XML (Mengel and Lezius, 2000) – popular for treebank encoding, or PAULA (Dipper, 2005) – a versatile format for multi-modal and multi-layered corpus encoding.[5] The differences pertain to details such as the assumed format of feature structures or the presence or absence of extra mechanisms, such as labelled edges (which can naturally be transduced into nodes when converting formats). We discuss the interrelations between the LAF family of ISO standards, Tiger-XML, PAULA, and the annotation schemas defined for the NCP in (Przepiórkowski, 2009; Przepiórkowski and Bański, 2010).[6] Przepiórkowski and Bański (2010) show that the particular TEI application for the NCP, a result of heavy customisation of the ultra-versatile toolkit that the TEI Guidelines offer, is (a) a concrete (“out-of-the-box”) solution subsuming the abstract GrAF, (b) isomorphic with Tiger-XML and PAULA, and often mirroring the devices used there, and (c) equipped with documentation trivially derivable from the literate-encoded ODD files (see below), (d) offering a homogeneous format for a variety of annotation layers and (e) offering well-tested metadata-encoding in the form of TEI headers that not only describe the source text and annotation documents but also (f) virtually link them, by being XIncluded into each of them. All annotation layers from the morphosyntactic layer upwards use the ISO/TEI feature structure representation (FSR) standard (ISO:24610-1). All in all, the NCP application of the TEI is offered for the encoders of complex corpora as a pragmatic solution that allows them to use a *homogeneous* set of well-documented schemas *interoperable* with the currently endorsed standards and best practices.

The above-mentioned ODD (“One Document Does it all”) files are the TEI’s recipe for what Bauman (2008) calls “literate encoding”, by reference to the literate programming paradigm (Knuth, 1984): TEI schemas are defined in TEI documents, with the typical TEI header and a standard text body with the addition of special elements that provide instructions for constructing schemes out of the content models and attribute classes offered by the TEI, cf. (Burnard and Rahtz, 2004). These files are then processed to derive schemas (such as DTD, RelaxNG, Schematron or XML Schema) and/or documentation in various formats. This provides for Helbig’s (2001) *communicability*, i.e. sharing uniform documentation across project members and with external entities.

The well-known TEI headers (due to their comprehensiveness and versatility used by many projects that do not use the TEI as such) provide for one of the aspects of sustainability, namely *discoverability*. The NCP headers record the history of the text (in extreme cases, also the entire headers of files that have been converted from the corpora created by the members of the NKJP Consortium) and the history of the annotation documents, classify the text, and provide all the standard information that can be useful in locating or querying the text. A single main corpus header provides information common to all files in the corpus and defines several taxonomies that the local headers use (examples of headers are provided at http://nlp.ipipan.waw.pl/TEI4NKJP/).

5. More on interoperability

In the previous section, we have addressed the issue of interoperability considered in terms of syntactic formats, i.e., from the point of view of what Witt *et al.* (2009) call

[4] See http://www.tei-c.org/Activities/Projects/ for an incomplete list of encoding projects using the TEI.

[5] In the case of Tiger-XML, the genealogy is different: it was created as an independent format and it is now being incorporated into ISO SynAF (ISO:24615). The NCP schema for syntactic annotation is isomorphic to SynAF/Tiger-XML.

[6] The TEI has re-incorporated the (X)CES proposals for corpus encoding (among others, stand-off annotation) and introduced its own schemes for referencing spans of characters and sequences of elements as extensions to the XPointer Framework (http://www.w3.org/TR/xptr-framework/). While the NCP demonstrates that the level of stand-off support in the TEI is sufficient for more technically-oriented users, there are still details that remain to be taken care of in order to ensure a greater level of the TEI’s user-friendliness in this regard. Some of them are discussed in Bański (2010).

static text-based LRs. In this section, we look at how the NCP copes with the semantic interoperability and move on to review the dynamic LRs (tools) offered by the project. Recall that semantic interoperability requires sharing a common data model and additionally, a common set of reference categories. In the context of ISO LAF (and LRs in general), one of the places offered to store reference categories is the ISOcat Data Category Registry (http://www.isocat.org/, cf. (Kemps-Snijders *et al.*, 2008)). Recently, as reported in (Patejuk and Przepiórkowski, 2010), the NKJP Tagset has been defined in the ISOcat (totalling in 85 Data Categories) and became the first public ISOcat definition of a complete tagset, available as a public Data Category Selection (keyword: nkjp). This testifies to the *semantic interoperability* potential of the NCP.

The main tools adapted for the NCP, Poliqarp (a search engine and a concordancer, cf. (Janus and Przepiórkowski, 2007))[7] and Anotatornia[8] are offered under the GNU General Public License (GPL). Anotatornia (Przepiórkowski and Murzynowski, forthcoming) is a tool for manual encoding of multi-level corpora (handling word-level and sentence-level segmentation as well as morphosyntax and word-sense disambiguation) that includes inter-annotator conflict-resolution mechanisms. These tools are the NCP's offer in the sphere of dynamic LRs. They are planned to be implemented in projects using TEI XML architecture with the data model based on that of the NCP, namely the Open-Content Text Corpus (OCTC, a multilingual SourceForge resource in the alpha stage of development, cf. (Bański and Wójtowicz, this volume)) and the Foreign Language Examination Corpus (a University of Warsaw Council for the Certification of Language Proficiency project, in the planning phase, cf. (Bański and Gozdawa-Gołębiowski, 2010)).

6. Conclusion

We have presented the National Corpus of Polish – a standards-compliant, scalable, and sustainable language resource with open-source tools designed to be flexible enough to interoperate with other resources of a similar type. The corpus contains a hierarchy of stand-off annotation levels, each of them is encoded in TEI XML, which satisfies the homogeneity requirement of Helbig (2001). Corpus documentation for each annotation layer can be derived from the appropriate ODD configuration files (Burnard and Rahtz, 2004), which fulfils the requirement of communicability.

The NCP demonstrates the usefulness of the TEI XML toolkit configured with an eye towards meeting the challenges that modern Language Resource producers and users face. These design choices have proven to be usable also for other resources of a similar general kind (the OCTC and the FLEC, mentioned above). We believe that this makes both the TEI – on the general plane, and the NCP – on the plane of applications, serious participants in the debate on the current state and future development in the sphere of Language Resources.

7. Acknowledgements

We are grateful to two anonymous LRSLM2010 reviewers for their helpful remarks. Additionally, Piotr Bański wishes to thank Victoria Arranz for her extraordinary patience.

8. References

Bański, P. (2010). Why TEI stand-off annotation doesn't quite work. Manuscript, University of Warsaw.

Bański, P. and Gozdawa-Gołębiowski, R. (2010). Foreign Language Examination Corpus for L2-Learning Studies. Submitted for the proceedings of the 3rd Workshop on Building and Using Comparable Corpora (BUCC), "Applications of Parallel and Comparable Corpora in Natural Language Engineering and the Humanities", 22 May 2010, Valletta, Malta.

Bański, P., Wójtowicz, B. (this volume). The Open-Content Text Corpus project. In proceedings of the LREC workshop on "Language Resources: From Storyboard to Sustainability and LR Lifecycle Management" (LRSLM2010), 23 May 2010, Valletta, Malta.

Bański, P. and Przepiórkowski, A. (2009). Stand-off TEI annotation: the case of the National Corpus of Polish. In Proceedings of the Third Linguistic Annotation Workshop (LAW III) at ACL-IJCNLP 2009, Singapore, pp. 64--67.

Bauman, S. (2008). Freedom to Constrain. In Balisage: The Markup Conference 2008, available at http://www.balisage.net/Proceedings/vol1/html/Bauman01/BalisageVol1-Bauman01.html.

Burnard, L., Rahtz, S. (2004). RelaxNG with Son of ODD. Presented at Extreme Markup Languages 2004, Montréal, Québec. Available from http://conferences.idealliance.org/extreme/html/2004/Burnard01/EML2004Burnard01.html

Dipper, S. (2005). XML-based stand-off representation and exploitation of multi-level linguistic annotation. In *Proceedings of Berliner XML Tage 2005* (BXML 2005). Berlin, pp. 39--50.

Głowińska, K., Przepiórkowski, A. (2010).The Design of Syntactic Annotation Levels in the National Corpus of Polish. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010*. Valletta, Malta.

Helbig, H. (2001). *Die semantische Struktur natürlicher Sprache. Wissensrepräsentation mit MultiNet*. Berlin: Springer.

Ide, N. (1998). Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora. *Proceedings of the First International Language Resources and Evaluation Conference*, Granada, Spain, pp. 463--470.

Ide, N. (2010). What does "interoperability" mean, anyway?. Keynote presentation delivered at ICGL-2010, City University of Hong Kong.

[7] http://poliqarp.sourceforge.net/

[8] http://nlp.ipipan.waw.pl/Anotatornia/

Ide, N., Bonhomme, P., Romary, L. (2000). XCES: An XML-based Standard for Linguistic Corpora. Proceedings of the Second Language Resources and Evaluation Conference (LREC), Athens, Greece, pp. 825--830.

Ide, N., Romary, L. (2007). Towards International Standards for Language Resources. In Dybkjaer, L., Hemsen, H., Minker, W. (eds), Evaluation of Text and Speech Systems, Springer, pages 263--284.

Ide, N., Suderman, K. (2007). GrAF: A Graph-based Format for Linguistic Annotations. Proceedings of the Linguistic Annotation Workshop, held in conjunction with ACL 2007, Prague, June 28-29, pp. 1--8.

Janus, D., Przepiórkowski, A. (2007). Poliqarp: An open source corpus indexer and search engine with syntactic extensions. In *Proceedings of the ACL 2007 Demo Session*. Prague. pp. 85--88.

Kemps-Snijders, M., Windhouwer, M.A., Wittenburg, P., Wright, S.E. (2008). ISOcat: Corralling Data Categories in the Wild. In European Language Resources Association (ELRA) (ed), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May 28-30, 2008.

Knuth, D.E. (1984). Literate programming. The Computer Journal (British Computer Society), 27(2), pp. 97--111.

Mengel, A., Lezius, W. (2000). An XML-based encoding format for syntactically annotated corpora. In *Proceedings of the Second Language Resources and Evaluation Conference* (LREC), Athens, Greece, pp. 121--126.

Młodzki, R., Przepiórkowski, A. (2009). The WSD Development Environment. In *Proceedings of the 4th Language & Technology Conference*. Poznań, Poland. pp. 185--189.

Nathan, D. (2006). Proficient, permanent, or pertinent: aiming for sustainability. In Barwick, L., Thieberger, N. (Eds.), *Sustainable Data from Digital Fieldwork*. The University of Sydney, pp. 57--68.

Patejuk, A., Przepiórkowski, A. (2010). ISOcat Definition of the National Corpus of Polish Tagset. In proceedings of the LREC 2010 workshop on “LRT Standards”. Valletta, Malta. ,

Przepiórkowski, A. (2009). TEI P5 as an XML Standard for Treebank Encoding. In: Passarotti, M., Przepiórkowski, A., Raynaud, S. Van Eynde, F. (eds), *Proceedings of of the Eighth International Workshop on Treebanks and Linguistic Theories* (TLT8), pp. 149--160.

Przepiórkowski, A., Bański, P. (2010). TEI P5 as a text encoding standard for multilevel corpus annotation. In Fang, A.C., Ide, N. and J. Webster (eds). *Language Resources and Global Interoperability. The Second International Conference on Global Interoperability for Language Resources (ICGL2010).* Hong Kong: City University of Hong Kong, pp. 133--142.

Przepiórkowski, A., Bański, P. (forthcoming). XML Text Interchange Format in the National Corpus of Polish. In S. Goźdź-Roszkowski (ed.) *The proceedings of Practical Applications in Language and Computers PALC 2009*, Frankfurt am Main: Peter Lang.

Przepiórkowski, A., Górski, R.L., Łaziński, M., Pęzik, P. (2010). Recent Developments in the National Corpus of Polish. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010*. Valletta, Malta.

Przepiórkowski, A., Murzynowski, G. (forthcoming). Manual annotation of the National Corpus of Polish with Anotatornia. In Goźdź-Roszkowski, S. (ed.) *The proceedings of Practical Applications in Language and Computers (PALC-2009)*. Frankfurt: Peter Lang.

Savary, A., Waszczuk, J., Przepiórkowski, A. (2010). Towards the Annotation of Named Entities in the National Corpus of Polish. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010*. Valletta, Malta .

Simons, G.F., Bird, S. (2008). Toward a global infrastructure for the sustainability of language resources. *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation: PACLIC 22*. pp. 87--100.

TEI Consortium (Eds.) (2010). TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 1.6.0. Last updated on February 12th 2010. TEI Consortium. http://www.tei-c.org/Guidelines/P5/

Thompson, H. S., McKelvie, D. (1997). Hyperlink semantics for standoff markup of read-only documents, Proceedings of SGML Europe. Available from http://www.ltg.ed.ac.uk/~ht/sgmleu97.html.

Witt, A., Heid, U., Sasaki, F., Sérasset, G. (2009). Multilingual language resources and interoperability. In *Language Resources and Evaluation*, vol. 43 :1, pp. 1--14. doi:10.1007/s10579-009-9088-x