Journal of the Text Encoding Initiative

Issue 3 | 2012
TEI and Linguistics

Editorial Introduction to the Third Issue

Piotr Bański, Eleonora Litta Modignani Picozzi and Andreas Witt

Publisher
TEI Consortium

Electronic version
URL: http://jtei.revues.org/544
DOI: 10.4000/jtei.544
ISSN: 2162-5603

Electronic reference
Piotr Bański, Eleonora Litta Modignani Picozzi, and Andreas Witt, « Editorial Introduction to the Third Issue », *Journal of the Text Encoding Initiative* [Online], Issue 3 | November 2012, Online since 05 November 2012, connection on 02 October 2016. URL : http://jtei.revues.org/544 ; DOI : 10.4000/jtei.544

Editorial Introduction to the Third Issue

Piotr Bański, Eleonora Litta Modignani Picozzi and Andreas Witt

1 Linguistics had a strong presence at the TEI's beginnings, being represented by names as significant as those of Nancy Ide, Donald E. Walker, and Antonio Zampolli. Linguistics was mentioned explicitly in the names of two of its three founding organizations: Association for Computers and the Humanities, Association for Computational Linguistics, and Association for Literary and Linguistic Computing. It was the main focus of one of the four initial committees (http://www.tei-c.org/Vault/AB/abj01.txt) and, within several years of the inception of the work on the TEI Guidelines, the British National Corpus clearly demonstrated the TEI's usefulness for encoding language resources.

2 While the TEI proved successful in annotating basic grammatical information in an in-line fashion, by the time the BNC was compiled there was a rapid development in corpus studies, directed not only at the volume of primary data but also at annotations that gradually began to provide information beyond part-of-speech categorization and lemmatization. Architectures were needed which would provide simple and fast deployment, describing exactly the information that was needed without the overhead of extra markup and using a flatter metadata structure. This is how specifications such as CES (primarily for morphosyntactic and alignment annotation) and TigerXML (for syntactic annotation, both hierarchical and relational) were developed and began to be adopted by the linguistic community.

3 The early 2000s saw rapid development of language resources encoded in CES (Corpus Encoding Standard) developed from TEI P3, and then XCES, as well as Tiger XML, both of which exceeded the TEI in their popularity within tightly focused linguistic circles. It should also be pointed out that, while the robust stand-off mechanisms of the TEI are still being refined, CES and then XCES provided basic reference mechanisms which proved extremely popular among corpus creators. Similar is the case of feature structure markup: while the ISO/TEI feature structure schema offers numerous ways to encode linguistic information (Witt et. al. 2009), in the absence of the feature structure validation mechanisms, corpus builders adopted much simpler solutions.

4 This state of affairs has gradually been changing: TEI P5, as a mature XML-based toolkit that supports all the newest XML technologies, once again could be an important player in the market of annotation standards (in the case of the TEI, a more precise phrase could be annotation standard toolkits) and has recently been applied to encode major linguistic enterprises such as the National Corpus of Polish, with its impressive stand-off architecture featuring a number of separate annotation layers (Przepiórkowski and Bański 2010).

5 The TEI special interest group for linguists (LingSIG), founded in 2010, has as its aim making the TEI even more competitive in the area of linguistic annotation frameworks, while maintaining close connections with the work performed at ISO TC37 SC4, the ISO committee devoted to the management of language resources.

6 At the time of writing, the SIG has met twice (at the TEI conferences in Zadar and in Würzburg) where a series of micropresentations were offered on various topics connecting the TEI and linguistics. It was also from the participants of the Würzburg meeting that most submissions for the present issue were received.

7 This issue begins with an overview of the current annotation standards landscape. “The TEI and Current Standards for Structuring Linguistic Data: An Overview,” by Maik Stührenberg, provides a remarkable summary of the most recent efforts to create international standards for the annotation of linguistic corpora, developed by the ISO technical committee for Terminology and other Language and Content Resources (ISO/TC 37). This article opens a window onto the world of standards creation, detailing the steps necessary for a set of protocols to become a standard, contrasting that with community discussion-based specifications such as the TEI Guidelines, and showing how the latter have been influential in the creation of *de facto* standards.

8 The second paper, “A TEI P5 Document Grammar for the IDS Text Model,” by corpus linguistics specialist Harald Lüngen and a veteran of both TEI and XML, C.M. Sperberg-McQueen, presents the process of making the legacy data of DeReKo (*Deutsches Referenzkorpus*, the largest archive of German written text, collected at the IDS Mannheim since 1964) compatible with the current version of TEI P5. The paper describes the steps taken to encode the corpus since the early 1990s through a detailed analysis of the way the IDS text model evolved to ultimately include the preparation of an ODD file which, in turn, documents the model.

9 Gerhard Budin, Stefan Majewski, and Karlheinz Mörth write about a similar effort in the area of dictionary encoding. Their paper describes the work of the Institute for Corpus Linguistics and Text Technology (ICLTT) of the Austrian Academy of Sciences in a number of projects involving both the digitisation of print dictionaries and the creation of new born-digital lexicographical data. The article explores how even within the restrictions imposed by the TEI dictionary module, an attentive customisation with an eye to interoperability with other standards and digital NLP tools makes TEI P5 a model that can be applied over a variety of digitisation projects. The article touches on issues of hierarchies, polyfunctionality of certain elements in the dictionary module, word-class information, and interoperability of the markup schema with other digital frameworks. The authors present the project’s experience in encoding morphosyntactic information, linguistic varieties and writing systems, etymology, semantics, and specific production metadata, ultimately proving the value of the customised TEI P5 dictionary module both

in the representation of digital dictionaries and the potential for use in NLP related applications.

10 "Consistent Modeling of Heterogeneous Lexical Structures" by Laurent Romary and Werner Wegstein highlights issues concerning the interoperability of a variety of data sources in lexical data modelling. This article starts by underlining the difficulties arising from building *ad hoc* data models from the TEI Guidelines' Dictionaries chapter, which inevitably leads to poor accessibility. The authors focus on lexical structures and propose a more generic methodology based on the concept of *crystals*, the smallest units in a construct that can help divide a document into regular chunks of information that can be processed more easily by external tools.

11 Michael Beißwenger, Maria Ermakova, Alexander Geyken, Lothar Lemnitzer, and Angelika Storrer present a novel application of TEI P5 in the description of computer-mediated communication. Their paper, "A TEI Schema for the Representation of Computer-mediated Communication," introduces an XML schema which provides a structure for the encoding of the structural units of communication in not only forums, blogs, and bulletin boards but also instant messaging, wikis, and twitter feeds, as well as the annotation of these units. The paper offers an interesting view on the processing of a new literary genre characterised by precise interaction features such as emoticons, interaction words, acronyms, and so on, and on the need for TEI P5 to cater for such forms of text.

12 Piotr Bański, Stefan Majewski, Maik Stührenberg, and Antonina Werthmann take on a more general issue relating to the social and infrastructural aspect of the SIG, and present a proposal for integrating a TEI markup exporter into the general-purpose citation manager Zotero. The paper provides a glimpse into the origins of the SIG's online presence and articulates a proposal for specific choices within TEI bibliographic elements to suggest a coherent and interchangeable way of sharing and maintaining bibliographic reference stores.

13 The guest editors of the volume wish to express their thanks to the authors and the reviewers, and acknowledge the work by the *Journal of the Text Encoding Initiative* regular editors, Susan Schreibman and Kevin Hawkins, in bringing the issue into uniform shape.

BIBLIOGRAPHY

Przepiórkowski, Adam and Piotr Bański. 2011. "XML Text Interchange Format in the National Corpus of Polish." In *Explorations across Languages and Corpora*, edited by Stanisław Góźdź-Roszkowski, 55–65. Frankfurt am Main: Peter Lang.

Witt, Andreas, Georg Rehm, Erhard Hinrichs, Timm Lehmberg, and Jens Stegmann. 2009. "SusTEInability of linguistic resources through feature structures." *Literary & Linguistic Computing* 24 (3): 363–372. doi:10.1093/llc/fqp024.