

XML-TEI-URS: using a TEI format for annotated linguistic resources

Loïc Grobol, Frédéric Landragin, Serge Heiden

▶ **To cite this version:**

Loïc Grobol, Frédéric Landragin, Serge Heiden. XML-TEI-URS: using a TEI format for annotated linguistic resources. CLARIN Annual Conference 2018, Oct 2018, Pisa, Italy. hal-01827563

HAL Id: hal-01827563
https://hal.science/hal-01827563

Submitted on 10 Sep 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

XML-TEI-URS: using a TEI format for annotated linguistic resources

Loïc Grobol
Lattice / ALMAnaCh
Paris, France
loic.grobol@ens.fr

Frédéric Landragin
Lattice
Paris, France
frederic.landragin@ens.fr

Serge Heiden
IHRIM
Lyon, France
slh@ens-lyon.fr

Abstract

This paper discusses XML-TEI-URS, a recently introduced TEI-compliant XML format for the annotation of referential phenomena in arbitrary corpora. We describe our experiments on using this format in different contexts, assess its perceived strengths and weaknesses, compare it with other similar efforts and suggest improvements to ease its use as a standard for the distribution of interoperable annotated linguistic resources.

1. Related works

XML-TEI-URS[1], introduced in (Grobol, Landragin, et al. 2017) is an annotation format inspired by the URS (Unit-Relation-Schema) metamodel developed for Glozz (Widlöcher and Mathet 2012) with a concrete serialization in TEI mark-up complying with the latest recommendations (TEI P5 v3.3.0). The original intent of this format was to provide a way to annotate reference phenomena, and particularly coreferences and anaphora, but it proved versatile enough for a larger class of annotations, as in (Grobol, Tellier, et al. 2018), where it is used for dependency syntax annotations. By design, it is not meant to be a ground-breaking new format, but rather a concrete realisation — within the limits of a standard serialization — of an abstract model proved to be sensible for coreference.

XML-TEI-URS is by no mean the first attempt at devising a general-purpose linguistic annotation format. There already exist several such formats, with wide range of uses, both in Corpus Linguistics and in Natural Language Processing, for example the tabular format used by BRAT (Stenetorp et al. 2012) or the XML-based formats used by GATE (Cunningham et al. 2013), MMAX2 (Müller and Strube 2006) or Glozz. But those formats are mostly tied to those specific annotation softwares — even when they express theoretically sound annotation models — and are susceptible to change along with their needs, with no guarantee of backward compatibility or notification of evolution. Consequently, they can only be thought of as *de facto* standards, whose use for perennial storage of linguistic resources could be problematic.

Conversely, as described in (Grobol, Landragin, et al. 2017), most of the annotated corpora for coreference use ad-hoc formats, that are usually well-suited to this single phenomenon, but do not support extension to other kind of annotations. The most common way to add other types of annotations (such as syntactic ones) in these resources is to use hybrid formats, such as in the tabular format used for the CoNLL-2012 corpus (Pradhan et al. 2012), which uses two different and incompatible types of parenthesized expressions for syntax and coreference annotations. One of the downsides of this approach is the data preparation overhead it imposes on the development NLP systems, a tedious and error-prone process, with scarce opportunities for reuse.

2. Experiments

Our experiments so far with XML-TEI-URS have been the following ones:

1. Porting the ANCOR corpus (Muzerelle et al. 2014) coreference annotations to XML-TEI-URS, first as a proof-of-concept for (Grobol, Landragin, et al. 2017).

1. This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/

2. Enriching ANCOR with syntactic annotations for (Grobol, Tellier, et al. 2018), which had us first convert it to a suitable input format for automatic parsers (Universal Dependencies CoNLL-U (Nivre et al. 2016)), then convert the resulting syntactic analysis back to XML-TEI-URS, with an additional annotation layer that describes the relations between syntax and coreference.

3. Integrating XML-TEI-URS in the URS annotation plugin of the TXM platform (Heiden 2010).

2.1. XML-TEI-URS for coreference: ANCOR

When it comes to file formats, ANCOR has a tumultuous story: it is composed of three different oral corpora, that were originally distributed in the native format of the transcription tool Transcriber (Barras et al. 1998). Its coreference annotations were then added in Glozz as if the data were raw texts, thus ignoring the existing XML structure (the consistency between the two layers was enforced manually), and finally integrated in it, using a non-standard ad-hoc format. This combination made the exploitation of this corpus tedious at best, and information-destructive in some cases (e.g. when entity mention crossed utterances borders). It was clear from the beginning of our work that there was a need for a better format — or at least one that was easier to use.

The initial conversion of ANCOR to XML-TEI-URS has actually been done at the same time as the definition of the format, which probably made the development of the necessary software tools more time-consuming than under other circumstances. Reflecting on that experience, we find that most of our difficulties came from the shortcomings of the original format, and from our efforts to enforce data consistency by correcting the errors that are inevitably present in any corpus of a significant size. All in all, the initial conversion took us no more than a few weeks, with some later refinements to meet unforeseen needs revealed by our actual use of the resulting corpus.

The resulting corpus is much easier to use than the original one, particularly thanks to the choice of completely stand-off annotations using a reference word segmentation (which are not mandatory for XML-TEI-URS, but heartily encouraged). The most welcomed advantage of this choice is that it allows to completely ignore the existence of annotations for preprocessing that does not take them into account (which is obviously harder with inline annotations) e.g. extracting the raw text of the corpus to run third-party tools on it is completely transparent.

The expressiveness of the TEI format also allowed annotations that were not possible in the original format, e.g. entity mentions spanning several utterances, or parallel and overlapping utterances.

2.2. XML-TEI-URS for syntax: ANCOR-AS

As stated in (Grobol, Tellier, et al. 2018), most automatic coreference detection systems use rich syntactic knowledge, which implies a need for corpora that hold both types of annotations. Existing corpora usually use one of three main strategies: use a ad-hoc hybrid format that incorporates the two types of annotations (as in CoNLL-2012), keep one version of the corpus for each type (as in the NER version of the French Treebank (Sagot et al. 2012)) or base one type on the other (as in the PCC, Polish Coreference Corpus (Ogrodniczuk et al. 2015)). In our context, none of these solutions was satisfactory. The most satisfactory would have been the option chosen for the PCC, but it requires mutually consistent annotation layers, which was not the case with automatic syntactic annotations.

Instead, we took advantage of the unobtrusive nature of standoff XML-TEI-URS annotations by totally ignoring existing coreference annotations at first when adding syntactic annotations, and only linking the two types of annotation in a third layer. The main obstacle in that process was that the word segmentation we used in the original version was not necessarily the same as the one given by the automatic parser. This issue was dealt with by adding correcting elements inspired by the then-current draft of (ISO 2017), that link between the surface forms of the raw corpus and the syntactic words used by the parser, in e.g. expansions (*du*→*de le* in French) and multi-word units.

Apart from this technical issue, the conversion between formats, from XML-TEI-URS to CoNLL-U and back was relatively straightforward, here again thanks to the use of reference to a word segmentation, for instance to clean up the parser inputs from easily detected disfluencies — thus improving its performances — while keeping them available in the raw text of the final resource.

That said, the final resource expresses the main drawback of the format: it is very heavy, far more than the corresponding CoNLL-U annotation, mostly because of our rather crude use of feature structure. Future versions of the resource will try to address this issue, most notably by a judicious use of MAF (ISO 2006) feature libraries, but a certain heaviness of TEI formats will always be unavoidable. In the meantime, we tried to mitigate this heaviness by keeping syntactic annotations in separated files, using the prefixed id facilities offered by the TEI to link them to the source files, which would have made sense in any case: since these syntactic annotations are not gold-standard, keeping them separated preserves the integrity of the manual coreference annotations of ANCOR.

2.3. Democrat and TXM platform

Since (Grobol, Landragin, et al. 2017), a progressive move towards using XML-TEI-URS as the format for the final version of the Democrat project corpus (Landragin 2016) is underway. In accordance, support for this format has been added to the URS annotation plugin developed in the context of this project for the TXM open-source platform. Full support for importing from and exporting to stand-off XML-TEI-URS is currently available, along with cross-corpus transfer of annotation between different corpora, as long as the tokens targeted by the annotations are present in both of them.

Integrating XML-TEI-URS to TXM was not too hard, thanks to the similarities with TXM internal format, which already used token-based stand-off annotations for lemmas and part-of-speech. The integration of XML-TEI-URS to TXM has been beneficial in reducing data duplication in version management, since several versions of a corpus can share the same outsourced annotation file as long as the token ids are constant. Conversely, several annotation sets can refer to the same corpus, which allows concurrent annotation by different annotators and keeping track of several versions of the same annotation set.

3. Conclusion and perspectives

Since its initial development, we have used XML-TEI-URS in several different contexts, and so far, it has lived up to our expectancies: it provides a standard, versatile and generally easy to use format for linguistic annotations. However, it comes with a certain heaviness that is probably linked to the TEI characteristics. This somewhat degrades human-reading experience for our prototype corpus, even though it had no impact on machine reading.

In the context of CLARIN, since the current guidelines advocate the use of TEI XML formats for textual data, we believe that XML-TEI-URS might serve as a basis — or at least an inspiration — for future linguistic resources with annotations going beyond the default set of attributes hardcoded into the current TEI guidelines (`lemma`, `pos` and friends). We are very much open to further developments or refinements of this format to better suit the needs of the community.

4. Acknowledgements

This work is part of the "Investissements d'Avenir" overseen by the French National Research Agency ANR-10-LABX-0083 (Labex EFL), and is also supported by the ANR DEMOCRAT (Describing and Modelling Reference Chains: Tools for Corpus Annotation and Automatic Processing) project ANR-15-CE38-0008.

References

[Barras et al. 1998] Claude Barras, Edouard Geoffrois, Zhibiao Wu, and Mark Liberman. 1998. Transcriber: a Free Tool for Segmenting, Labeling and Transcribing Speech. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA). Granada, España, May 1998.

[Cunningham et al. 2013] Hamish Cunningham, Valentin Tablan, Angus Roberts, and Kalina Bontcheva. 2013. Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. *PLOS Computational Biology*, 9.2, Feb. 2013: 1–16.

[Grobol, Landragin, et al. 2017] Loïc Grobol, Frédéric Landragin, and Serge Heiden. 2017. Interoperable annotation of (co)references in the Democrat project. In Harry Bunt, editor, *Thirteenth Joint ISO-ACL Workshop on Interoperable Semantic Annotation*. ACL Special Interest Group on Computational Se-

mantics (SIGSEM) and ISO TC 37/SC 4 (Language Resources) WG 2. Montpellier, France, Sept. 2017.

[Grobol, Tellier, et al. 2018] Loïc Grobol, Isabelle Tellier, Éric De La Clergerie, Marco Dinarelli, and Frédéric Landragin. 2018. ANCOR-AS: Enriching the ANCOR Corpus with Syntactic Annotations. In *LREC 2018 - 11th edition of the Language Resources and Evaluation Conference*. Miyazaki, Japan, May 2018.

[Heiden 2010] Serge Heiden. 2010. The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In *24th Pacific Asia Conference on Language, Information and Computation (PACLIC)*, pages 389–398. Institute for Digital Enhancement of Cognitive Development, Waseda University. Sendai, Japan, Nov. 2010.

[ISO 2006] ISO/TC 37/SC 4. 2006. *ISO 24610-1:2006 Language resource management – Feature structures – Part 1: Feature structure representation*. Reference. Geneva, CH: International Organization for Standardization, Apr. 2006.

[ISO 2017] ISO/TC 37/SC 4/WG 2. 2017. *ISO AWI 24617-9 Language resource management – Part 9 Reference Annotation Framework (RAF)*. Reference. Geneva, CH: International Organization for Standardization, 2017.

[Landragin 2016] Frédéric Landragin. 2016. Description, modélisation et détection automatique des chaînes de référence (DEMOCRAT). *Bulletin de l'AFIA*, 92, 2016: 11–15.

[Müller and Strube 2006] Christoph Müller and Michael Strube. 2006. Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, Kurt Kohn, and Joybrato Mukherjee, editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang. Frankfurt a.M., Germany, 2006.

[Muzerelle et al. 2014] Judith Muzerelle et al. 2014. ANCOR Centre, a Large Free Spoken French Coreference Corpus: Description of the Resource and Reliability Measures. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA). Reykjavík, Ísland, May 2014.

[Nivre et al. 2016] Joakim Nivre et al. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In Nicoletta Calzolari et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA). Portorož, Slovenia, May 23–28, 2016.

[Ogrodniczuk et al. 2015] Maciej Ogrodniczuk, Katarzyna Głowińska, Mateusz Kopeć, Agata Savary, and Magdalena Zawisławska. 2015. *Coreference in Polish: Annotation, Resolution and Evaluation*. Walter De Gruyter, 2015.

[Pradhan et al. 2012] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. In *Proceedings of the joint Conference on Empirical Methods in Natural Language Processing (EMNLP) and Computational Natural Language Learning (CoNLL)*, pages 1–40. CoNLL '12. Association for Computational Linguistics. Jeju, Korea, 2012.

[Sagot et al. 2012] Benoît Sagot, Marion Richard, and Rosa Stern. 2012. Annotation référentielle du Corpus Arboré de Paris 7 en entités nommées. In Georges Antoniadis, Hervé Blanchon, and Gilles Sérasset, editors, *Traitement Automatique des Langues Naturelles (TALN)*. Volume 2 - TALN. Actes de la conférence conjointe JEP-TALN-RECITAL 2012. Grenoble, France, June 2012.

[Stenetorp et al. 2012] Pontus Stenetorp et al. 2012. BRAT: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. Association for Computational Linguistics. Avignon, France, Apr. 2012.

[TEI P5 v3.3.0] TEI consortium, editor. 2018. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 3.3.0. TEI P5: Guidelines for Electronic Text Encoding and Interchange. Jan. 31, 2018. URL: `http://www.tei-c.org/Guidelines/P5` (visited on 04/24/2018).

[Widlöcher and Mathet 2012] Antoine Widlöcher and Yann Mathet. 2012. The Glozz Platform: A Corpus Annotation and Mining Tool. In *Proceedings of the 2012 ACM Symposium on Document Engineering*, pages 171–180. DocEng '12. ACM. Paris, France, 2012.