Interoperable annotation of (co)references in the Democrat project

Loïc Grobol[1, 2], Frédéric Landragin[1], Serge Heiden[3]

(1) Lattice, CNRS, ENS Paris, Université Sorbonne Nouvelle,
PSL Research University, USPC, 1 rue Maurice Arnoux, 92120 Montrouge, France
(2) ALMAnaCH, Inria, 2 rue Simone Iff, 75589 Paris, France
(3) IHRIM, ENS Lyon, CNRS, University of Lyon, 15 parvis René Descartes 69342 Lyon, France

{loic.grobol, frederic.landragin}@ens.fr ; slh@ens-lyon.fr

Abstract

This paper proposes XML-TEI-URS, a generic TEI-based format for the annotation of coreferences in arbitrary corpora. This proposal is made in the context of Democrat, a French *Agence Nationale de la Recherche* project that aims to produce a large corpus of written French with coreference annotations, in an attempt to design a corpus that is usable both by humans and automated tools and as compatible as possible with future concurrent annotations.

1 Introduction

In this paper we propose XML-TEI-URS, a TEI-compliant format for coreference annotation inspired by Glozz' URS (Unit-Relation-Schema) metamodel (Widlöcher and Mathet 2012). Our main goal is to provide a standard that is easy to process for automated tools, compatible with other types of annotations and stays as human-readable as possible, in order to improve the interoperability of annotated corpora.

This proposal is formulated in the context of the Democrat project (Landragin 2016), which aims to produce a large corpus of written French with coreference annotations. It also takes inspiration from the MC4 projet (Mélanie-Becquet and Landragin 2014) —which has served as a proof-of-concept for the Democrat project—, the ANCOR (Muzerelle et al. 2014) and PCC (Polish Coreference Corpus (Ogrodniczuk et al. 2015)) corpora, which fulfilled similar objectives, respectively for oral French and written Polish and previous standardisation initiatives, such as Bruneseaux and Romary (1997).

We first give a rough description of the existing formats for coreference-annotated corpora, then go on to look at the tools available in the TEI for coreference annotation, and finally describe how we plan to implement coreference annotation in the Democrat corpus and in a TEI-compliant version of the ANCOR corpus.

2 Context

Historically, the first corpus with coreference annotations was the MUC corpus (Grishman and Sundheim 1996) released in 1995 in the occasion of the 6th Message Understanding Conference and annotated using simple inline SGML. Its natural successor, ACE (Doddington et al. 2004), which was dedicated to the related Entity Detection and Tracking task and released in 2003 used a similar format, but with stand-off annotations to replace some of the inline ones used in the MUC corpus.

In 1999, in the context of the MATE project Poesio, Bruneseaux, and Romary (1999) proposed a generic scheme for annotating coreference in the MATE workbench (Isard et al. 2000), both in terms of representation format and annotation conventions. The MATE project and its follow up (Poesio 2004) inspired the format of several subsequent corpora, most notably the GNOME (Poesio 2000) and AnCora (Taulé, Martí, and Recasens 2008) corpora and directly influenced the markup scheme used by the MMAX2 annotation format in which several corpora have been annotated. Among them were ARRAU (Poesio and Artstein 2008), LiveMemories (Rodríguez et al. 2010) and EPEC-KORREF (Soraluze et al. 2012). To our knowledge, this was the first effort of standardization of coreference annotations. These formats were XML-based, using elements similar —but not identical— to those proposed by the TEI for stand-off annotation. While being similar, they were not generally compatible, and not necessarily easy to reuse.

A major event for coreference-annotated corpora was the release in 2007 of the first edition of the OntoNotes corpus (S. S. Pradhan et al. 2007), followed by several other editions. It is to this day the largest coreference-annotated corpus, with nearly 3M words in three languages –Arabic, English and Chinese. Its format was inspired from those of the MUC and ACE corpora, with inline SGML annotations for coreference (and other types of annotation, most notably for parsing in independent files and formats) and with a methodology inspired in part by the MATE guidelines. The use of the OntoNotes corpus in the CoNLL-2011 and CoNLL-2012 (S. Pradhan et al. 2012) shared tasks made it the standard evaluation ground for coreference detection systems, but in that case, the compatibility of the corpus and its ease of reuse in other contexts was far from being evident.

An example of a coreference corpus that followed a TEI-format is the PCC, developed by Ogrodniczuk et al. (2015). Its format is an extension of the one used by its base, the NKJP corpus (Adam Przepiórkowski and Łaziński 2008), and coreference annotations are implemented as stand-off annotations in separate files.

Regarding French, for now, the only publicly available[1] large-scale coreference-annotated corpus is ANCOR (Muzerelle et al. 2014), a 418k words corpus of transcribed oral. This transcription imposed the base of its format, as the source corpora were in the Transcriber's (Barras et al. 1998) TRS format. As for the coreference annotations, they started as stand-off Glozz annotations in separate files, that were later integrated into the original Transcriber files. The resulting corpus has thus an non-standard format that is importable neither in Transcriber nor in Glozz and can be cumbersome to parse for other applications. Another corpus of importance for our reflexions is MC4, a diachronic corpus of written French, annotated for the MC4 project as a proof-of-concept. Its format, XML-TEI-Analec, is the base of our reflexion here, presenting TEI-compliant stand-off coreference annotations in self-contained files.

The Democrat project, started in 2015, aims at creating and using a large-scale coreference-annotated diachronic corpus of written French by 2019. To that aim, it plans to leverage the experience gathered in the similar but smaller-scaled MC4 project. The manual annotation of this corpus is planned to be done with Analec (Landragin, Poibeau, and Victorri 2012) at first, while phasing it out for appropriate tools implemented for the TXM platform (Heiden 2010). For now, this annotation procedure yields native Analec files, with optional conversion to the XML-TEI-Analec format.

3 Linguistic annotations and coreference in the TEI

The description of an annotation in any kind of representation is made of two parts: a way of marking the element (markable) that is being annotated (unitizing) and a way of stating the annotations you are giving about it (categorizing).

For coreference (as for most linguistic annotations), the markables are usually contiguous spans of text. Unitizing is thus easily done using TEI's `<span>` element (or possibly `<seg>` as in the PCC). The main point of divergence is on how to specify the boundaries of those spans. Following Bański et al. (2016), we can identify three main pointing mechanisms in the TEI for stand-off annotations:

- Offset-based mechanisms such as TEI pointers (see Cayless (2013) for details and perspectives)
  - Used by the PCC and NKJP corpora through TEI's `string-range` and separate files.
  - Does not require source text alteration, but has to refer to a specific version of it.
- In-source `<anchor>` or similar markers
  - Used by Analec to denote markable (e.g. mentions) boundaries.
  - Requires minimal source text alteration and does not limit concurrent annotations.
  - Considerable source text clutter (particularly with concurrent annotations).
- Referring to `<w>` words/tokens
  - Used by TXM to mark words (tokens) for lemmatisation and POS-tagging.
  - Requires source text alteration and the non-trivial choice of a segmentation
  - Does not ease character-level boundaries encoding.
  - Considerably eases the aggregation of e.g. syntactic and coreference annotations.

As for categorizing, the TEI offers the `<fs>` structure, that can support a variety of annotations, including recursive structures, multivalued features, free-text features… This mostly solves the issue of annotating mention features — such as definiteness or syntactical type — but also cluster-level features — such as associativity or agreement — that have no support in the source text.

Finally, it is of course crucial for coreference annotation to be able to denote relations between mentions. There are usually two types of relations that have to be annotated: coreference relation between pairs of entity,

[1] Tutin et al. (2000) presents another large-scale corpus with anaphoric links annotations, which is not publicly available. It uses a close derivative of the MATE format.

which can be directed (from a source to a target) in the case of anaphora ; and coreference of a set of mentions (usually called a *coreference chain*). While there is no explicit way of specifying this kind of relation in the TEI, one can be designed with existing elements. Since mentions are already annotated, these relations can simply be annotated as references to them. There are mechanisms for this usage in the TEI, under the *linking* and *aggregating* categories (TEI consortium 2016). Analec uses `<join>` elements to that purpose, which is arguably non-conform with the stated purpose of these elements. The PCC corpus uses the more generic `<ptr>`, which is more relevant, but makes the association between elements of a coreference chain harder, as it then requires to look for siblings of a given `<ptr>`. A third possibility exists in the form of the `<link>` element, which "defines an association [...] among elements". This last alternative seems to be the most standard way of marking both URS relations and URS schema in the TEI, the only inconvenient being the necessity of relying on the order of `xml:id` in the `target` attribute to mark the direction of directed relations. It is also coherent with the recommendations of (Bruneseaux and Romary 1997).

4 XML-TEI-URS: a format for coreference annotation

In this section, we describe the XML-TEI-URS format: a lightweight format for the annotation of coreferences. This format is heavily inspired from the XML-TEI-Analec format described in Mélanie-Becquet and Landragin (2014), with an effort to improve its compliance with the spirit of the TEI guidelines and its integration in various types of TEI-formatted documents. The core of the format is the implementation in TEI elements of the URS metamodel introduced by Widlöcher and Mathet (2012).

- Mentions are *units* in the Glozz sense, i.e. contiguous[2] spans of the source text. They are represented by `<span>` elements with `type="unit"`.
- Coreference relations between mentions are represented by `<link>` elements with `type="relation"` whose `target` attributes are, in that order, the `xml:id` of the source mention and the `xml:id` of the target mention.
- Coreference chains are represented by `<link>` elements with `type="schema"` whose `target` attributes are the `xml:id` of their member mentions, in no particular order.
- All of these might point to a `<fs>` using `ana` attributes to encode further linguistic annotations.

While they may seem redundant in the context of coreference, relations and schema are both needed for the annotation of more sophisticated concepts — for example bridging anaphora — that are naturally represented as directed relations between coreferences chains. This justify their inclusion here, however close the concepts and their concrete representations may seem.

Note that none of theses requirements specify where those elements should be located. This is due to the lack of a dedicated place in TEI documents for stand-off annotations. In our experiments with ANCOR, we follow Romary (2017) and use the `<standOff>` element, which has not yet made its way into the official TEI guideline. If full compliance with the current TEI guidelines is required, however we see two main strategies: either place coreference annotation in the `<back>` of the document (as in Analec), possibly in a dedicated `<div>` or place them in the `<body>` of an external file (as in the PCC corpus). Either way, we encourage the use of `<spanGrp>` and `<linkGrp>` to avoid redundancy in the annotation. In the same way, to ensure a better hierarchy of annotations, we suggest that `<fs>` elements be grouped together in dedicated `<div>` elements, thus separating different types of annotations. These do not allow factoring out informations as do e.g. `<spanGrp>`, but they should provide easier information retrieval for corpus users. See example 1 for an application to the ANCOR corpus.

Example 1: XML-TEI-URS for the ANCOR corpus

```
<text>
<body>
    <div type="section" xml:id="s2">
        <timeline>
            <when absolute="3.531" xml:id="t2.0"/>
            [...]
            <when absolute="25.924" xml:id="t7.0"/>
            [...]
        </timeline>
        <u start="#t2.0" who="#spk1" xml:id="u2" end="#t2.1">
```

[2] Contiguity is actually not required for `<span>`, so this could be easily extended to non-contiguous units.

```
            [...]
        </u>
        [...]
        <u start="#t7.0" who="#spk2" xml:id="u7" end="#t7.19">
            [...]
            <w xml:id="u7-w20">en</w>
            <w xml:id="u7-w21">octobre</w>
            <w xml:id="u7-w22">mille</w>
            <w xml:id="u7-w23">neuf</w>
            <w xml:id="u7-w24">cent</w>
            <w xml:id="u7-w25">soixante</w>
            <anchor synch="#t7.6" type="time"/>
            [...]
            <w xml:id="u7-w36">à</w>
            <w xml:id="u7-w37">ce</w>
            <w xml:id="u7-w38">moment-là</w>
            <w xml:id="u7-w39">après</w>
            [...]
        </u>
        [...]
    </div>
    [...]
</body>
</text>
<standOff>
    <div type="coreference">
        <spanGrp type="unit">
            [...]
            <span ana="#m2713-fs" from="#u7-w21" to="#u1-w25" xml:id="m2713"/>
            <span ana="#m2714-fs" from="#u7-w37" to="#u1-w38" xml:id="m2714"/>
            [...]
        </spanGrp>
        <div type="unit-fs">
            [...]
            <fs xml:id="#m2713-fs">
                <f name="type">
                  <string>DATE</string>
                </f>
                [...]
            </fs>
            [...]
        </div>
        <linkGrp type="relation">
            [...]
            <link ana="#r212-fs" target="#m2713 #m2714" xml:id="r212"/>
            [...]
        </linkGrp>
        [...]
        <linkGrp type="schema">
            [...]
            <link target="#m2713 #m2714 #m2731" xml:id="s150"/>
            [...]
        </linkGrp>
    </div>
</standOff>
```

The format described above does not impose a particular pointing mechanism from those listed in section 3. However, in the specific context of the Democrat project, mentions boundaries may not occur inside of words, which relieves us from the burden of supporting character-level annotation. In addition to this, the TXM platform that is planned to support the final corpus already allows orthographic tokenisation as a support for its own internal annotations. This considerations lead us to recommend the use of the third solution, namely annotating mentions as spans anchored to <w> elements as shown in example 1.

As a support for the development and test of our proposed format, we used it to convert the ANCOR corpus to a fully TEI-compliant format. This involved both converting its original transcription-related annotation to their TEI equivalents and converting its coreference annotations to the format described above, example 1 shows an example of this. This shows the capacity of our format to adapt to different corpus paradigms, and bodes well for further applications, including to documents with more complex TEI representations. It also eases the access to the annotations of ANCOR for the automated annotation tools we are developing in the Democrat project (which was part of our initial motivation) compared to both the original ANCOR format and MC4's XML-TEI-Analec format.

It should also be noted that this format, while designed for coreference annotation, is not restricted to this application. Indeed, Glozz' URS metamodel has been shown to be suitable for a variety of applications such as analysis of opinion or discourse structures, which gives us hopes for the adoption of our format for other kinds of annotations, thus increasing the general interoperability of annotated corpora.

5 Conclusion and perspectives

In this paper, we presented a quick overview of the formats used for existing coreference-annotated corpora. While some efforts of standardization exist, it is our beliefs that those do not fulfill our goals of simplicity and interoperability for the annotation of the Democrat corpus. The XML-TEI-URS format we propose is TEI-compliant and plays well with other types of linguistic annotations, including transcription annotations for oral corpora. At the same time, it offers some leeway, particularly regarding pointing mechanisms, in order to accommodate different exigences for corpus makers and users. This proposal is also intentionally agnostic regarding the theoretical framework used for the actual annotation of (co)reference. In that sense, it is complementary to of e.g. the future recommendations of ISO (2017) and potentially any other paradigm.

We worked within the limits of the TEI, which already provides suitable tools for reference annotation. However, we deplore the lack of dedicated mechanisms to attach stand-off annotations to documents, and it is our hope that the integration of the <standOff> element in the TEI guidelines would fulfill that lack.

Our next steps should be the release of both an explicit TEI XML schema for our format, and the release of the whole ANCOR corpus in this format as a proof-of-concept. In the context of the Democrat project, we will also have to refine the actual features we use for the annotation of coreference phenomena, features that could then make their way into our recommendations for the annotation of coreference in other projects.

6 Acknowledgements

This work is part of the "Investissements d'Avenir" overseen by the French National Research Agency ANR-10-LABX-0083 (Labex EFL).

This work has been supported by the ANR DEMOCRAT (Description et modélisation des chaînes de référence: outils pour l'annotation de corpus et le traitement automatique) project ANR-15-CE38-0008.

References

Adam Przepiórkowski, Barbara Lewandowska-Tomaszyk, Rafał L. Górski, and Marek Łaziński. 2008. "Towards the National Corpus of Polish." In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08).* Http://www.lrec-conf.org/proceedings/lrec2008/. Marrakech, Morroco: European Language Resources Association (ELRA), May.

Bański, Piotr, Bertrand Gaiffe, Patrice Lopez, Simon Meoni, Laurent Romary, Thomas Schmidt, Peter Stadler, and Andreas Witt. 2016. *Wake up, standOff!* TEI Conference 2016. Wien, Austria, September.

Barras, Claude, 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).* Granada, España: European Language Resources Association (ELRA), May.

Bruneseaux, Florence, and Laurent Romary. 1997. "Codage des références et coréférences dans les DHM." In *ACH-ALLC'97.* Kingston, Canada.

Cayless, Hugh A. 2013. "Rebooting TEI Pointers." *Journal of the Text Encoding Initiative* (6).

Doddington, George, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. “The Automatic Content Extraction (ACE) Program : Tasks, Data, and Evaluation.” In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-2004).* ACL Anthology Identifier: L04-1011. Lisboa, Portugal: European Language Resources Association (ELRA), May.

Grishman, Ralph, and Beth Sundheim. 1996. “Message Understanding Conference-6: A Brief History.” In *Proceedings of the 16th Conference on Computational Linguistics - Volume 1,* 466–471. COLING ’96. København, Danmark: Association for Computational Linguistics.

Heiden, Serge. 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),* 389–398. Sendai, Japan: Institute for Digital Enhancement of Cognitive Development, Waseda University, November.

Isard, Amy, David McKelvie, Andreas Mengel, and Morten Baun Møller. 2000. “The MATE Workbench: A Tool for Annotating XML Corpora.” In *Content-Based Multimedia Information Access - Volume 1,* 411–425. RIAO ’00. Paris, France: Centre de hautes études internationales d’informatique documentaire.

ISO/TC 37/SC 4/WG 2. 2017. *ISO AWI 24617-9 Language resource management – Part 9 Semantic annotation framework (SemAF).* Reference. Geneva, CH: International Organization for Standardization.

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

Landragin, Frédéric, Thierry Poibeau, and Bernard Victorri. 2012. “ANALEC: a New Tool for the Dynamic Annotation of Textual Data.” In *International Conference on Language Resources and Evaluation (LREC 2012),* edited by European Language Resources Association (ELRA), 357–362. İstanbul, Türkiye, May.

Mélanie-Becquet, Frédérique, and Frédéric Landragin. 2014. “Linguistique outillée pour l’étude des chaînes de référence questions méthodologiques et solutions techniques.” *Langages,* no. 195 (September): 117–137.

Muzerelle, Judith, Anaïs Lefeuvre, Emmanuel Schang, Jean-Yves Antoine, Aurore Pelletier, Denis Maurel, Iris Eshkol, and Jeanne Villaneau. 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).* Reykjavík, Ísland: European Language Resources Association (ELRA), May.

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

Poesio, Massimo. 2000. “Annotating a Corpus to Develop and Evaluate Discourse Entity Realization Algorithms: Issues and Preliminary Results.” In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC,* 211–218. Athenes, Greece.

———. 2004. “The MATE/GNOME Proposals for Anaphoric Annotation, Revisited.” In *Proceedings of the 5th SIGDIAL Workshop,* 154–162. Boston, MA, USA, April.

Poesio, Massimo, and Ron Artstein. 2008. “Anaphoric Annotation in the ARRAU Corpus.” In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC-08).* Marrakech, Morroco: European Language Resources Association (ELRA), May.

Poesio, Massimo, Florence Bruneseaux, and Laurent Romary. 1999. “The MATE meta-scheme for coreference in dialogues in multiple languages.” In *ACL’99 Workshop Towards Standards and Tools for Discourse Tagging,* 65–74. College Parc, MD, USA, June.

Pradhan, Sameer S., Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2007. “OntoNotes: A Unified Relational Semantic Representation.” In *Proceedings of the International Conference on Semantic Computing,* 517–526. ICSC ’07. Washington, DC, USA: IEEE Computer Society.

Pradhan, Sameer, 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),* 1–40. CoNLL ’12. Jeju, Korea: Association for Computational Linguistics.

Rodríguez, Kepa Joseba, Francesca Delogu, Yannick Versley, Egon W. Stemle, and Massimo Poesio. 2010. “Anaphoric Annotation of Wikipedia and Blogs in the Live Memories Corpus.” In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10).* Valletta, Malta, May.

Romary, Laurent. 2017. “stdfSpec : A proposal for a stand-off element for the TEI Guidelines.” July 7. `https://github.com/laurentromary/stdfSpec`.

Soraluze, Ander, Olatz Arregi, Xabier Arregi, Klara Ceberio, and Arantza Díaz de Ilarraza. 2012. “Mention detection: First steps in the development of a Basque coreference resolution system.” In *Proceedings of KONVENS 2012,* edited by Jeremy Jancsary, 128–136. Main track: oral presentations. Wien, Austria: ÖGAI, September.

Taulé, Mariona, M. Antònia Martí, and Marta Recasens. 2008. “AnCora: Multilevel Annotated Corpora for Catalan and Spanish.” In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC-08).* ACL Anthology Identifier: L08-1222. Marrakech, Morroco: European Language Resources Association (ELRA), May.

TEI consortium, ed. 2016. “16 Linking, Segmentation, and Alignment.” TEI P5: Guidelines for Electronic Text Encoding and Interchange. December 15. Accessed July 7, 2017. `http://www.tei-c.org/release/doc/tei-p5-doc/en/html/SA.html`.

Tutin, Agnès, François Trouilleux, Catherine Clouzot, Éric Gaussier, Annie Zaenen, Stéphanie Rayot, and Georges Antoniadis. 2000. “Annotating a large corpus with anaphoric links.” In *Third International Conference on Discourse Anaphora and Anaphor Resolution (DAARC2000),* 2. United Kingdom.

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