

# XML-TEI-URS: USING A TEI FORMAT FOR ANNOTATED LINGUISTIC RESOURCES

Loïc Grobol<sup>1,2</sup>, Frédéric Landragin<sup>1</sup> & Serge Heiden<sup>3</sup>

(1) Lattice (CNRS, ENS Paris, Université Sorbonne Nouvelle, PSL Research University, USPC)

(2) ALMAnaCH (Inria) (3) IHRIM (ENS Lyon, CNRS, University of Lyon)

## Transcription of speech in ANCOR

```
<div type="section" xml:id="s2">
  <timeline>
    <when absolute="3.531" xml:id="t2.0"/>
    [...]
  </timeline>
  <u start="#t7.0" who="#spk2" xml:id="u7">
    end="#t7.19"
    [...]
    <w xml:id="u7-w76">au</w>
    <w xml:id="u7-w77">moment</w>
    <w xml:id="u7-w78">où</w>
    <w xml:id="u7-w79">je</w>
    <w xml:id="u7-w80">me</w>
    <w xml:id="u7-w81">suis</w>
    <w xml:id="u7-w82">marié</w>
    <w xml:id="u7-w83">en</w>
    <w xml:id="u7-w84">juillet</w>
    <w xml:id="u7-w85">soixante-sept</w>
  </u>
</div>
```

- Porting from Transcriber's (Barras et al. 1998) TRS internal format
- No loss of expressivity
- Much more precise metadata and semantic markup
- Useful extensions for overlapping utterances (that had to be reconstituted)
- Allows for a reference segmentation using `<tei:w>`

## Coreference annotations in ANCOR

```
<standOff>
  <annotation type="coreference">
    <spanGrp type="unit" subtype="mention">
      <span from="#u7-w76" to="#u7-w77" xml:id="m31" ana="#m31-fs"/>
      <span from="#u7-w84" to="#u7-w85" xml:id="m32" ana="#m32-fs"/>
    [...]
  </spanGrp>
  <linkGrp type="relation" subtype="coreference">
    <link target="#m31 #m32" xml:id="r20"/>
  [...]
  </linkGrp>
  <linkGrp type="schema" subtype="chain">
    <link target="#m31 #m32 #m40" xml:id="c12"/>
  [...]
  </linkGrp>
</annotation>
</standOff>
```

- The original motivation for the format creation (Grobo, Landragin, et al. 2017)
- Ported from Glozz (Widlöcher et al. 2012) annotations interspersed in the original TRS files
  - Our version is far easier to parse
- Based on the URS annotation metamodel implemented in TEI XML
- Uses experimental stand-off TEI markup (Romary 2017)
  - Allows complex parallel annotation layers
  - Easier parsing of the main content and linking with the annotations
  - Possibility of annotations crossing discourse (utterance) boundaries, discontinuous annotations...
- Similar, but not identical to the upcoming ISO RAF serialization (ISO 2018)

## Perspectives

- Improve compacity, reduce verbosity (via StandOff evolutions)
- Improve the compatibility with RAF serialization (as soon as the standard is published)
- Develop further visualization/editor facilities and integrate with existing ones
- Publish ODD/Schematron specifications and documentation and propagate

## Dependency syntax in ANCOR-AS

```
<standOff>
  <annotation type="syntax">
    <div type="multiword-token">
      <expan xml:id="tree10-w6-7" n="6-7" corresp="#u7-w76">
        <w xml:id="u7-w76.1">à</w>
        <w xml:id="u7-w76.2">le</w>
      </expan>
    </div>
    <div type="tree" xml:id="tree10">
      <spanGrp type="unit" subtype="word">
        [...]
        <span target="#u7-w76.1" n="6" xml:id="tree10-w6"/>
        <span target="#u7-w76.2" n="7" xml:id="tree10-w7"/>
        <span target="#u7-w77" n="8" xml:id="tree10-w8"/>
        [...]
      </spanGrp>
      <linkGrp type="relation" subtype="dependency">
        [...]
        <link target="#tree10-w8 #tree10-w6" xml:id="tree10-d6"/>
        <link target="#tree10-w8 #tree10-w7" xml:id="tree10-d7"/>
        <link target="#tree10-w3 #tree10-w8" xml:id="tree10-d8"/>
        [...]
      </linkGrp>
    [...]
  </annotation>
</standOff>
```

- Extends URS to include dependency syntax in the UD (Nivre et al. 2016) formalism (Grobo, Tellier, et al. 2018)
  - Take into account diverging segmentations
  - Allows for separate annotation of merged words as in (fr) *du* → *de le*
  - Full typing/feature structure for dependencies (overkill for UD)
- Synchronization of syntax and coreference annotations, even in problematic cases e.g. mentions not fitting to a subtree
- Main issue: considerably heavier files, not necessarily easily human-readable

## Democrat and TXM platform

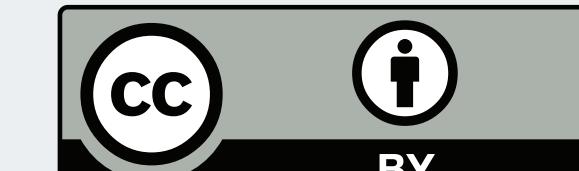
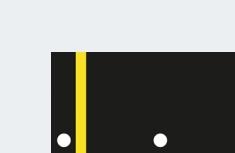
et le **povre Caillette** demeuroit  
là, et ne disoit mot: Car il n'avoit point  
d'autre apprehension, sinon qu'il pensoit  
estre confiné là pour toute sa vie. Il passe **un des**  
**Seigneurs de court**, qui le voit ainsi en conseil avec ce  
piller, qui le fait incontinent desgager de là: s'enquerant  
bien expressement qui avoit fait cela, et qui l'a  
mis là? Que voulez **vous**, un **sot** l'a mis là, un **sot** l'a  
mis. Quand on disoit. Ce ont esté les **baees**. Caillette

- Motivation: have the two big coreference corpora —ANCOR (Muzerelle et al. 2013) and DEMOCRAT (Landragin 2016)— in similar formats
- Development of a general purpose annotation tool in TXM (Heiden 2010) suitable for DEMOCRAT
- TXM already uses an XML-TEI format for its internal purposes
  - Heavy ad-hoc metadata that would not make sense outside of TXM
  - The similarities and standard interface made the integration of XML-TEI-URS relatively straightforward
- Reduces duplication in annotation versioning and parallel annotations

## References

- Barras, Claude et al. (May 1998). "Transcriber: a Free Tool for Segmenting, Labeling and Transcribing Speech". In: *Proceedings of the First International Conference on Language Resources and Evaluation (LREC)*.
- Grobo, Loïc, Frédéric Landragin, et al. (Sept. 2017). "Interoperable annotation of (co)references in the Democrat project". In: *Thirteenth Joint ISO-ACL Workshop on Interoperable Semantic Annotation*.
- ACL Special Interest Group on Computational Semantics (SIGSEM) and ISO TC 37/SC 4 (Language Resources) WG 2.
- Grobo, Loïc, Isabelle Tellier, et al. (May 2018). "ANCOR-AS: Enriching the ANCOR Corpus with Syntactic Annotations". In: *LREC 2018 - 11th edition of the Language Resources and Evaluation Conference*.
- Heiden, Serge (Nov. 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)*.
- ISO/TC 37/SC 4/WG 2 (2018). *ISO AWI 24617-9: Language resource management – Part 9 Reference Annotation Framework (RAF)*. Reference. International Organization for Standardization.
- Landragin, Frédéric (2016). "Description, modélisation et détection automatique des chaînes de référence (DEMOCRAT)". In: *Bulletin de l'AFAI 92*.
- Muzerelle, Judith et al. (June 2013). "ANCOR, premier corpus de français parlé d'enverge annoté en corréférence et distribué librement". In: *TALN'2013, 20e conférence sur le Traitement Automatique des Langues Naturelles*.
- Nivre, Joakim et al. (May 23–28, 2016). "Universal Dependencies v1: A Multilingual Treebank Collection". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.
- Romary, Laurent (July 7, 2017). *stdfSpec: A proposal for a stand-off element for the TEI Guidelines*. URL: <https://github.com/laurientromary/stdfSpec>.
- Widlöcher, Antoine et al. (2012). "The Glozz Platform: A Corpus Annotation and Mining Tool". In: *Proceedings of the 2012 ACM Symposium on Document Engineering*.

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 (Describing and Modelling Reference Chains: Tools for Corpus Annotation and Automatic Processing) project ANR-15-CE38-0008.



This work is licensed under a Creative Commons "Attribution 4.0 International" license.