XML-TEI-URS: USING A TEI FORMAT FOR ANNOTATED LINGUISTIC RESOURCES
Loïc Grobol1,2, Frédéric Landragin1 & Serge Heiden3
(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

- Publish ODD/Schematron specifications and documentation and propagate
- Improve compacity, reduce verbosity (via StandOff evolutions)
- Similar, but not identical to the upcoming ISO RAF serialization (ISO 2018)
- Based on the URS annotation metamodel implemented in TEI XML
- Ported from Glozz (Widlöcher et al. 2012) annotations interspersed in the original files
- The original motivation for the format creation (Grobol, Landragin, et al. 2017)
- Allows for a reference segmentation using <tei:p>

Coreference annotations in ANCOR

- The original motivation for the format creation (Grobol, 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)

Democrat and TXM platform

- 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


Grobol, Loïc, Isabelle Tellier, et al. (May 2018). "ANCOR-AS: Enriching the ANCOR Corpus with Syntactic Annotations".


Heiden, Serge (2010). "Developing a General Purpose Annotation Tool in TXM".

Muzerelle, Judith et al. (June 2013). "ANCOR, premier corpus de français parlé d’envergure annoté en coréférence et distribué librement".
In: Bulletin de l’AFIA. 316-319. 92.

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