View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by INRIA a CCSD electronic archive server

DGfS CL

Aria Adli[1], Eric Engel[1], Laurent Romary[2], Fahime Same[1]
1. University of Cologne, 2. INRIA

A stand-off XML-TEI representation of reference annotation

In this poster, we present an XML-TEI conformant stand-off representation of reference in discourse, building on the seminal work carried out in the MATE project (Poesio, Bruneseaux & Romary 1999) and the earlier proposal on a reference annotation framework in Salmon-Alt & Romary (2005).

We make a three-way distinction between *markables* (the referring expressions), *discourse entities* (referents in the textual or extra-textual world), and *links* (relations that hold between referents, e.g., part-whole). Our approach differs from previous suggestions in that (i) inherent properties of the referent itself (e.g., animacy) are disentangled from the expressions used to refer to that referent, which is both conceptually clearer and results in faster annotation especially on longer stretches of text, (ii) existing annotations from other layers such as morphosyntax are cleanly separated from the annotation of reference, but can be combined in queries and (iii) our proposal is integrated into the larger structure of existing TEI-ISO standards, such as MAF and SynAF, thereby allowing for compatibility with existing TEI-encoded corpora and data sustainability. This work is part of an ongoing discussion on the ISO project 24617-9 Language resource management — Semantic annotation framework — Part 9: Reference.

The workflow of adding reference annotations to an existing corpus will be demonstrated with ongoing work on the *sgs* corpus (http://www.sgscorpus.com), which already contains morphosyntactic annotations in a stand-off TEI format. In particular, we will show how reference information added in MMAX2 (Müller & Strube 2006) can be transformed to the proposed format and how it can be combined with existing annotations to perform increasingly complex queries. We give concrete examples from ongoing work in the SFB 1252 (subprojects C01 and INF), where this representation of reference is the backbone for the annotation of topic chains in dialogue data and for queries of topics in various grammatical constructions.

ISO:24611 (2005). Language resource management – Morphosyntactic annotation framework (MAF). ISO/CD 24611, ISO TC 37/SC 4 document N225 of 2005-10-15.

ISO:24615 (2009). Language resource management – Syntactic annotation framework (SynAF). ISO/CD 24615, ISO TC 37/SC 4 document N421 of 2009-01-30.

Müller, C. & M. Strube. 2006. Multi-level annotation of linguistic data with MMAX2. In S. Braun, K. Kohn & J. Mukherjee (eds.), *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, 197–214. Frankfurt a.M.: Peter Lang.

Poesio, M., F. Bruneseaux & L. 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, United States. https://hal.inria.fr/inria-00525171.

Salmon-Alt, S. & L. Romary. 2005. The Reference Annotation Framework: A case for semantic content representation. In H. Bunt (ed.), *IWCS-6*, Tilburg, Netherlands: ACL SIGSEM. https://hal.inria.fr/inria-00489935.