Recent Initiatives towards New Standards for Language Resources

Gottfried Herzog, Ulrich Heid, Thorsten Trippel, Piotr Bański, Laurent Romary, Thomas Schmidt, Andreas Witt, Kerstin Eckart

Recent Initiatives towards New Standards for Language Resources

Gottfried Herzog[1], Ulrich Heid[2], Thorsten Trippel[3], Piotr Bański[4], Laurent Romary[5], Thomas Schmidt[4], Andreas Witt[4], Kerstin Eckart[6]

[1]Deutsches Institut für Normung e. V., Berlin,
gottfried.herzog@din.de
[2]Universität Hildesheim, [3]Universität Tübingen,
[4]Institut für Deutsche Sprache, Mannheim, [5]Inria, [6]Universität Stuttgart

1 Introduction

This poster is aimed at providing an overview of three ongoing initiatives towards language resource (LR) standards coordinated and initiated by the German mirror group of ISO TC 37/SC 4[1] within DIN[2] (Deutsches Institut für Normung):

- ISOTiger, an XML serialization of proposals for the syntactic annotation of text corpora;
- "Transcription of spoken language", a set of guidelines for transcribing spoken utterances;
- "Corpus Query Lingua Franca", a meta-standard for the comparison of the formal properties of corpus query languages.

Coordinated by German experts, these upcoming international standards[3] are all part of initiatives to standardize data formats and procedures for language resources internationally. The present poster is intended not only to inform about the ongoing work, but also to initiate a discussion with additional experts to reflect the interests of the community.

Standards for LRs in the framework of ISO TC 37 cover several types of resources (text corpora, lexicons, terminology collections). Actors in computational linguistics and language technology cooperate and thus need to exchange data and technologies using comparable methods and formats, cf. (Eckart and Heid, 2014). Most of the proposed standards are guidelines on a meta-level, describing properties of representation formats, instead of prescribing a format. Examples of these are the *Lexical markup framework* (LMF, ISO 24613:2008), the *Terminological markup framework* (TMF, ISO 16642:2003) and the *Linguistic annotation framework* (LAF, ISO 24612:2012) for lexical or terminological entry representation and for the representation of annotated corpora, respectively.[4]

In corpus annotation, more specific standards have been developed, for linguistic annotation at the levels of morphosyntax (MAF, ISO 24611:2012) and syntax (constituency and dependency, SynAF, ISO 24615-1:2014), as well as for certain aspects of semantic annotation, such as e.g. the annotation of temporal expressions (SemAF-Time, ISO-TimeML, ISO 24617-1:2012).

The proposed paper describes work towards standards which will be integrated into the existing standards portfolio: ISOTiger (ISO/DIS 24615-2) is building on top of SynAF; the standard for transcription of audio- or video-recorded spoken interactions (*Transcription of spoken language*, ISO/CD 24624) fills a gap in the domain of the preparation of spoken corpora; and the third proposal (CQLF, *Corpus Query Lingua Franca*, ISO/CD 24623-1) targets properties of tools for querying corpora.

2 ISOTiger

ISOTiger is an XML serialization of the SynAF meta-model. For this serialization, TIGER-XML (König et al., 2003), a widely applied corpus encoding format, which originated form the German TiGer project (Brants et al., 2004), was enhanced to meet the SynAF requirements for a generic exchange format for syntactic annotations. This includes independence from a specific theoretical orientation or annotation scheme: there shouldn't e.g. be any preferences whether the annotation consists of constituency trees or dependency graphs, or whether the encoded information results from a deep or a shallow analysis. ISOTiger fits in with existing serializations for other annotation

[1]International Organization for Standardization, Technical Committee 37, Subcommittee 4: http://www.iso.org/iso/home/standards_development/list_of_iso_technical_committees/iso_technical_committee.htm?commid=297592

[2]http://www.nat.din.de/cmd?level=tpl-untergremium-home&committeeid=54739043&languageid=de&bcrumblevel=2&subcommitteeid=63074672

[3]They are now progressing to the level of "Draft International Standard", the last feedback option before publication.

[4]The appendix contains a reference list of all standards discussed in this paper.

Proceedings of the Int. Conference of the German Society for Computational Linguistics and Language Technology, pages 154–156, University of Duisburg-Essen, Germany, Sep 30–Oct 2 2015.

layers: morpho-syntactic annotations encoded according to a MAF serialization naturally constitute the leaves of ISOTiger-encoded syntax trees in a standoff annotation. Moreover, we are discussing to use the full power of feature structures, cf. (FSR, ISO 24610-1:2006), in ISOTiger, cf. (Bosch et al., 2014). Similar to LAF and MAF, SynAF separates the structure of the annotations from the semantics of the annotation categories, thus it is possible within ISOTiger to link elements of tagsets to external data categories describing their semantics, cf. (ISO 12620:2009).

3 Transcription of spoken language

The standard on *Transcription of spoken language* is motivated, similar to the corpus representation standards, by the need to compare, interchange and possibly combine transcriptions of spoken language; this also concerns tool environments for the creation, editing, publication and exploration (e.g. query) of transcribed data. The standard is based on a comparative study of state of the art tools and their formats, and it is compatible with widely used transcription systems. The standard is being developed in cooperation with TEI proposals in the field, cf. (Schmidt, 2011).
It addresses metadata (briefly, as more standards proposals for this domain are available in CMDI (ISO 24622-1:2015) and from the TEI), as well as the macro- and microstructure of transcriptions. The macrostructure involves the timeline, as well as single or grouped utterances and elements outside utterances (e.g. `<pause>` and `<incident>` items).
The microstructure proposals deal in depth with the annotation of tokens, pauses, audible or visible non-speech events, punctuation, as well as units above and below the level of utterances. It also includes recommendations concerning the handling of uncertain cases, alternatives, incomprehensible or omitted passages. The appendices contain an ODD specification and a fully encoded example.

4 CQLF: Corpus Query Lingua Franca

CQLF proposes a standardized metamodel for classifying the data models underlying different corpus query languages (=QLs). It distinguishes three levels of QL complexity and thereby opens up a space of properties of QLs. The first level covers query systems for linear annotation, i.e. plain text or simple annotations to segments. Level 2 in addition involves complex annotations, either hierarchical (as in constituent structures) or dependency-like. Level 3 adds concurrent annotations, i.e. cases where a given phenomenon has been annotated in multiple ways which may overlap, be intersecting or even in conflict. The current part I will be complemented by an ontology of QL features, guidelines for the development of customized QLs (part II), as well as an analysis of QLs for multimodal and parallel corpora (part III). The specification provides general guidelines on a rudimentary classification of QLs, together with several examples in the annex.

The poster will present key elements of the three initiatives; all of them are thoroughly documented, and interested parties should not hesitate to contact the German experts on the DIN committees with comments and suggestions to the proposals. Work on all three proposals will continue in 2015 and early 2016, and at least for CQLF over a longer time frame.

References

Sonja Bosch, Kerstin Eckart, Gertrud Faaß, Ulrich Heid, Kiyong Lee, Antonio Pareja-Lora, Laurette Pretorius, Laurent Romary, Andreas Witt, Amir Zeldes, and Florian Zipser. 2014. From <tiger2> to ISOTiger – Community Driven Developments for Syntax Annotation in SynAF. In Verena Henrich, Erhard Hinrichs, Daniël de Kok, Petya Osenova, and Adam Przepiórkowski, editors, *Proceedings of the Workshop on Treebanks and Linguistic Theories*.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. TIGER: Linguistic Interpretation of a German Corpus. *Research on Language and Computation*, 2(4):597–620.

Kerstin Eckart and Ulrich Heid. 2014. Resource interoperability revisited. In Josef Ruppenhofer and Gertrud Faaß, editors, *Proceedings of KONVENS*, volume 1, pages 116–126. Universität Hildesheim.

Esther König, Wolfgang Lezius, and Holger Voormann, 2003. *TIGERSearch 2.1 User's Manual. Chapter V*. IMS, Universität Stuttgart.

Thomas Schmidt. 2011. A TEI-based approach to standardising spoken language transcription. *Journal of the Text Encoding Initiative*, Issue 1. [Online], `http://jtei.revues.org/142`.

Appendix A. Reference list of mentioned standards and standard proposals

ISO 16642:2003	Computer applications in terminology – Terminological markup framework
ISO 24610-1:2006	Language resource management – Feature structures – Part 1: Feature structure representation
ISO 24611:2012	Language resource management – Morpho-syntactic annotation framework (MAF)
ISO 24612:2012	Language resource management – Linguistic annotation framework (LAF)
ISO 24613:2008	Language resource management – Lexical markup framework (LMF)
ISO 24615-1:2014	Language resource management – Syntactic annotation framework (SynAF) – Part 1: Syntactic model
ISO/DIS 24615-2	Language resource management – Syntactic annotation framework (SynAF) – Part 2: XML serialization (ISOTiger)
ISO 24617-1:2012	Language resource management – Semantic annotation framework (SemAF) – Part 1: Time and events (SemAF-Time, ISO-TimeML)
ISO 24622-1:2015	Language resource management – Component Metadata Infrastructure (CMDI) – Part 1: The Component Metadata Model
ISO/CD 24623-1	Language resource management – Corpus Query Lingua Franca (CQLF) – Part 1: Metamodel
ISO/CD 24624	Language resource management – Transcription of spoken language