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]DIN Deutsches Institut für Normung e. V., Berlin,
[2]Universität Hildesheim, [3]Universität Tübingen, [4]Institut für Deutsche Sprache, Mannheim, [5]Inria, [6]Universität Stuttgart

E-mail: gottfried.herzog@din.de

Standardization Work

- Who?
  - Experts from industry, academia and administrations
  - Experts are nominated – based on expertise and interest
  - ISO committee TC 37/SC 4 and national mirror committees,
    e.g. for Germany: *Normenausschuss NA-105-00-06 AA Sprachressourcen*
    DIN – Deutsches Institut für Normung e.V.
- How?
  - Stepwise procedures:
    Proposals – working drafts – (draft) international standards
  - Consensus-based: drafting – commenting – ballot
  - National standards organizations provide infrastructure

Excerpt of the list of standards and standard proposals by ISO TC 37/SC 4

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

by ISO TC 37/SC 3

ISO 12620:2009	Terminology and other language and content resources – Specification of data categories and management of a Data Category Registry for language resources
ISO 16642:2003	Computer applications in terminology – Terminological markup framework

ISOTiger

An XML-serialization for the SynAF metamodel.

SynAF – the syntactic annotation framework

- Generic exchange format for syntactic annotations
- Implements independence from specific theoretical orientation or annotation scheme: no preferences in terms of, e.g., constituency or dependency structures, deep or shallow analysis, etc.
- Is part of a larger set of corpus representation formats, for individual annotation layers, cf. MAF, SemAF, etc.
- Implements the separation of structure and content: makes use of a DCR according to ISO 12620

ISOTiger

Example on handout

- Based on TIGER-XML, an existing and widely used format, rather than 'inventing' a completely new format. **[König et al. 2003]**
- Modified wrt TIGER-XML: to represent dependency structures, to relate annotations to a DCR, to allow for different node and edge types, etc.
- Full power of feature structures – for a discussion see: **[Bosch et al. 2014]**

Transcription of spoken language

A representation format to compare, interchange and combine orthography-based transcriptions of spoken language.
The standard is developed in cooperation with TEI proposals in the field. **[Schmidt 2011]**

Based on

- State of the art tools and formats for creating, editing, publishing and querying transcribed data
- Widely used transcription systems

Encoded components

Example on handout

- Metadata: based on TEI header
- Macrostructure: timeline, single and grouped utterances, elements outside utterances (e.g. <pause> and <incident>)
- Microstructure:
  - annotations of tokens, pauses, audible or visible non-speech events, punctuation, units above and below the level of utterances
  - recommendations for uncertain cases, alternatives, incomprehensive or omitted passages

CQLF: Corpus Query Lingua Franca

A metamodel for classifying corpus query languages with respect to their (formal) properties (part I of the standard).

Levels of complexity

- Level 1 (linear): plain text search and segment annotations
- Level 2 (complex): annotated hierarchical structures and dependencies
- Level 3 (concurrent): multiple annotations from the same layer: overlapping, intersecting or conflicting

Overview of possible search patterns

		Search pattern				
		Plain text	Simple annotations	Complex annotations: Hierarchies	Complex annotations: Dependencies	Concurrent annotations
Linear	L1 – a)	+				
	L1 – b)		+			
	L1 – c)	+	+			
Complex	L2 – a)		+	+		
	L2 – b)		+		+	
	L2 – c)		+	+	+	
	L2 – d)	+	+	+		
	L2 – e)	+	+		+	
	L2 – f)	+	+	+	+	
Concurrent	L3 – a)		+			+
	L3 – b)		+	+		+
	L3 – c)		+		+	+
	L3 – d)		+	+	+	+
	L3 – e)	+	+			+
	L3 – f)	+	+	+		+
	L3 – g)	+	+		+	+
	L3 – h)	+	+	+	+	+

Complementary parts to come

- Part II: ontology of query language features, guidelines for the development of customized query languages on a specific level
- Part III: criteria for query languages for multimodal and parallel corpora

Bosch et al. 2014 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*.

Brants et al. 2004 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.

König et al. 2003 Esther König, Wolfgang Lezius, and Holger Voormann. 2003. *TIGERSearch 2.1 User's Manual. Chapter V - The TIGER-XML treebank encoding format*. IMS, Universität Stuttgart.

Schmidt 2011 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.