Standards for Language Resources

Overview and Use

Over the past 10 years, standardisation initiatives within W3C, ISO and the TEI have produced a suite of standards that directly impact the way language resources are and will be developed. This trend has in turn led to significant enhancements in the interoperability of tools for annotating and managing language resources, and to a better understanding of the issues involved in ensuring that such resources remain usable in the long term. This tutorial provides an overview of several key standards for language resources and demonstrates their use for annotation, description, and analysis of language resources.

Tutorial Overview

9.00-9.30 Background and core standards
Laurent Romary

9.30-10.30 Encoding primary resources – TEI, TEI-ODD
Lou Burnard

11.00-12.00 ISO TC37 SC4 standards: Feature structures, Morphological Annotation Format (MAF), Syntactic Annotation Format (SynAF), and Word Segmentation
Eric de la Clergerie and Andreas Witt

12.00-13.00 Time and Dialogue Acts
Harry Bunt and James Pustejovsky

- Lunch break -

14.30-16.00 Linguistic Annotation Framework (LAF)
Nancy Ide and Keith Suderman

16.30-18.00 ISOCat for data category references
Sue Ellen Wright and Menzo Windhouwer

18.00-19.00 ISOCat and GrAF Demonstrations

Session Descriptions

International standards for language resources – Background and core standards

This session overviews of existing standardization activities (ISO, TEI, W3C) that impact the domain of language resources and language technology, together with an outline of reference standards for our field (e.g., XML, together with ISO 10646/Unicode, ISO 639 series/language codes, ISO 15924/language scripts).
Presenter: Laurent Romary; INRIA & HUB-IDSL

Encoding primary resources – TEI, TEI-ODD

The TEI (Text Encoding Initiative), initiated in 1987 as a joint effort involving experts from the computational linguistics and humanities computing communities, has over the last twenty years evolved into the reference standard for textual information of all kinds. Its modularity, flexibility, and hospitality to variation distinguish it from many other standards endeavours.
Presenter: Lou Burnard

Basic annotation: Feature structures, MAF, SynAF, WordSeg

This section is dedicated to a comprehensive presentation of the basic annotation-building blocs provided by ISO for morphosyntactic and syntactic annotation. Starting with a background presentation of the Feature Structure Representation Standard (jointly developed with the TEI consortium), several examples will show how various levels of complexity (segmentation ambiguities, multi-word unit ambiguity, constituent vs. dependency analysis) can be dealt with in MAF and SynAF, as well as how

this relates to the ongoing work on word segmentation in the context of WordSeg (part 1 and 2) for, in particular, CJK languages.
Presenter: Eric de la Clergerie and Andreas Witt

Semantic annotation: Time and Dialogue acts

This session provides an overview of the recent advances within ISO in the domain of semantic annotation, with a focus on the coherent principles that have lead to the establishment of the multi-part standard SemAF (Semantic Annotation Framework). A more detailed example will illustrate how the TimeML format has been moved to an ISO standard, as well as an update on the current work on dialogue act annotation.
Presenter: James Pustejovsky, Harry Bunt

Lunch break

Linguistic Annotation Framework

This session overviews the Linguistic Annotation Framework (LAF) developed within ISO committee TC 37/SC 4. It shows usage cases involving the creation of annotated corpus data and how to map user annotation formats to GrAF (Graph Annotation Format), the XML serialization of the LAF pivot format. It also demonstrates the requirements for exchange via the LAF pivot, including creation of headers and use of the ISOCat data category registry for annotation content. We will also demonstrate the use of tools and APIs for accessing and manipulating annotations in GrAF format.

Presenters: Nancy Ide, Keith Suderman

Using ISOCat to manage data category references for linguistic annotation.

This section will focus on presenting the underlying principles and actual features of ISOcat, the linguistic concept database that has been set within ISO committee TC 37 to provide reference semantics for all types of annotation features (*data categories*) and values. This database can be used in conjunction with the other ISO standards to provide maximal interoperability across implementation (e.g. allowing the comparison of tagsets for morphosyntactic annotation with MAF).
Presenters: Menzo Windhouwer (demo) & Sue Ellen Wright (guidelines)