LIRICS

Linguistic Infrastructure for Interoperable Resources and Systems

- **Presentation of Actual Work on Standardization for Language Resources at DIN (NAT AA6)**

- **Thorsten Trippel, Thierry Declerck, Ulrich Heid**

Overview:

- Standards for Language Resources
- Word Segmentation
- Morpho-Syntactic Annotation Framework (MAF)
- Linguistic Annotation Framework (LAF)
- Lexical Markup Framework (LMF)
- Feature Structure Representation (FSR)
- Syntactic Annotation Framework (SynAF) – new ISO Work Item intreoduced by LIRICS.
- Data Categories (DatCats)
- Conclusions

Standards for Language Resources

- Heterogeneous Resources:
  - Linguistics
  - Language Technologies
  - Translation
  - Lexicon development
- Idiosyncratic data formats
- Problems:
  - Data exchange
  - Re-usability

Standards for Language Resources (2)

- Possible Solution:
  - Standardization of Language Resources:
    - Best practice
    - Portability
- ISO TC 37/SC4 (Language Resources)
  - DIN NAT (AA 6) in Germany and other national bodies
  - LIRICS, promoting the whole line of standards, with new Work Items (SynAF)

Word Segmentation

- A must in automated language processing
- Problem:
  - Not always by blanks
  - Treatment of coumpounds
  - Evaluation of tools/processing strategy missing
- Goal: A Metamodel for Segmentation (for the time being in MAF)
  - Word property of Multi-Word-Expressions
  - Linguistic rules

Morpho-Syntactic Annotation Framework (MAF)

- Goal: Unitary codification of morpho-syntactic annotation
- Content of MAF:
  - Segmentation
  - Content description of the annotation
- State: At an advanced level in the ISO procedure (DIS)

Linguistic Annotation Framework (LAF)

- Goal: Unitary base for the annotation of Linguistic Data
  - XML based, incl. Semantic Web representation languages
  - Stressing on higher level of annotation
- Content of LAF: A generic Data Format
  - Based on results of ISLE/EAGLES, TEI
- State: At the beginning of the ISO Procedure (WD)

Lexical Mark-Up Framework (LMF)

- Goal: Exchange Format for lexical databases Stressing on higher level of annotation
  - Similar to ISO 12200 (Martif)
  - Including dictionaries
- Content: Unitary Model for representation of dictionaries and (computational) lexicons
- State: At the beginning of the ISO Procedure (WD)

Feature Structure Representation (FSR)

- Goal: Collect all kind of feature representations schemes used in Language Technology
- Content: Unitary Syntax for feature structures, on the base of TEI
- State: CD submitted to DIS approval

Syntactic Annotation Framework (SynAF)

- Goal: Propose a meta model for syntactic annotation
- Content: Constituency and Dependency structures
- State: Submitted as a NewWork Item by the LIRICS Consortium

Data Categories (DatCats)

- Goal: Propose (neutral) definition of data categories that subsume various encodings and formats used in different systems.
- Content: Analog to ISO 12600 for terminology
- Lirics will propose such list for Lexicons, MAF, SynAF, SemAF (to come after the project lifetime)
- Discussion: open list of DatCats (Control?)

Conclusions

- LIRICS: Embedding of ISO activities in a EC Project (eContent framework).
- Strong commitments of DIN NAT AA6 (and other AA)