RELISH meets LOEWE

Frankfurt, 10-10-2011

Jost Gippert

RELISH

- Rendering Endangered languages Lexicons Interoperable through Standards Harmonization
- DFG/NEH Bilateral Digital Humanities Program, 2009-2012
- Frankfurt (University), Nijmegen (MPI), Chicago (Eastern Michigan University)

The RELISH mission

- When a lexicon constitutes the only record of a dying or already extinct language, it can contribute unique linguistic and cultural information to our store of scientific knowledge. And making it interoperable with other lexical data becomes a critical research priority. However, despite the support accorded to initiatives to develop digital standards for language documentation within both the US and Germany, there still exist major barriers to lexicon interoperability. The most significant barrier is that standards-setting bodies have arrived at different standards for format and markup on the two sides of the Atlantic. Additionally, within each national community, divergences exist in lexicon format and markup, in part because field linguists have hitherto relied on software which does not offer the linguist adequate support in choosing structural or linguistic categories.

Comparison of lexical entries

Javanese	**ojot** ng/di*(i) 1 to smooth [a bamboo stalk] with a knife. 2 to cut [a bamboo stalk] into shorter lengths	Two sense definitions grouped into a single entry.
Orokolo	**para**[1] (var. **pāra**) [pore] [T. *pōle*] *n.* a kind of mangrove (*Rhizophora*). **para**[2] [T. *poti*] *n.* club. (Mt. 26:55) **para**[3] [T. *fara*] *n.* an oar; (fr. MT **bara** 'oar'). **para koa** [T. *fara toaf*] *v.t.* to ply oar, to row.	One entry for each sense definition.
Urdu	s. بساهن *bisāhan*, s. بساهند *bisāhind*, } (बास Smell) s. f. Fetidness, stink.	Two entries with a single sense definition.

- In the 55 lexicons which they analyzed, Bird and Bell found 55 different types of lexical entries. And they examined primarily printed dictionaries (although they privileged dictionaries based on primary data). Had they examined unpublished lexicons of endangered languages produced by field linguists, they would certainly have discovered an even greater variety of structure and terminology.

Frankfurt: Caucasian Languages (DoBeS project ECLinG)

- Udi (ISO 639-3: udi): East-Caucasian (Lezgi) language of Azerbaijan / Georgia, with less than 3000 speakers; Toolbox
  - Gukasyan 1974: Udi-Russian-Azeri Dictionary
- Batsbi / Tsova-Tush (ISO 639-3: bbl): East-Caucasian (Nakh) Language of Georgia, with less than 1500 speakers
  - Kadagidze-Kadagidze 1984: Batsbi-Georgian-Russian Dictionary

Frankfurt: Caucasian Languages (DoBeS project ECLinG)

- Svan (ISO 639-3: sva): South Caucasian language of North West Georgia with less than 15,000 speakers
  - Topuria / Kaldani 2000: Svan-Georgian Dictionary and others
- Megrelian (ISO 639-3: xmf): South Caucasian language of West Georgia with ca. 500,000 speakers
  - Kajaia 2001-2004: Megrelian-Georgian Dictionary

LOEWE
“Digital Humanities”

- State of Hesse “LOEWE” program
- 2011-2013
- Partners: Goethe University Frankfurt, Technical University Darmstadt, Goethe Museum Frankfurt (Freies Deutsches Hochstift)
- www.dhhe.de

Project areas

- Historical corpora
- Modern language corpora
- Archive, corpus, edition
- Multimodal corpus management

Historical Corpora

- Detection and verification of cross-textual relationships in historical corpora
  - Translated texts
  - Parallel corpora
    - Bible texts
  - Diachronical stratification and linguistic change
  - Genetic affinity of languages
  - Historical semantics

Modern Language Corpora

- Corpus analysis
  - Text as a product
    - Machine learning and automatical corpus analysis
  - Text as an instance of the linguistic system
    - Non-canonical grammatical constructions
  - Text as process
    - Texts produced collaboratively (Web 2.0)

Archive, Corpus, Edition

- Digital redefinition of philological tasks
  - Collection vs. Archive vs. Edition
  - “Critical Apparatuses”
    - Avesta
  - Visualization → Online Editions
    - Faust, Hofmannsthal
  - Image-Text-Relations
    - Manuscript tradition of texts

Multimodal Corpus Management

- Development of annotation models
  - Multimodal multilevel annotation
  - Data models for statistical analysis
  - Flexible annotation models
  - Text technological integration

RELISH meets LOEWE

- Structuring and analysis of linguistic data
- Interoperability
- Multimodality
- Ontologies
- and much more...