From Cooperative GIS to Language Monitor: Coordinating Documentation Activities and Sharing Data

Oliver Streiter

National University of Kaohsiung, Taiwan,

ostreiter@nuk.edu.tw

Problem Description: Language Documentataion out of Sync

- Coordinating documentation projects
- Identify documentation needs (common grid?)
- Sharing information: access, informants, transcription systems
- Fast updates: Population, speakers, legal changes
- Involving communities: Native speakers in diaspora,
- Integrating partial, suboptimal documentation

Problem Solution: A Web-based Wiki-fied Language Monitor

- Leaving a trace of a project (past, present, future)
- Identify documentation needs, e.g automatic comparison of available data
- Sharing information among projects (maps, photos, addresses, word-lists)
- Instantanious online updates
- Involving communuties as partners (open projects, feedback, more data)
- Upgrading, elaborating, commenting on partial, suboptimal documentation

Do we just need a http://wiki.ethnologue.org ?

- Yes, we need a http://wiki (interface!).
- Yes, we need an .org
- But something different from Ethnologue is needed
  - Structured Data
  - Geo-referential Metadata
  - Model-based Data
  - Theory-based Data

Structured Data, no Prose !

a n g u a g e s

ข้อมูล

Argobba

A language of Ethiopia

ISO 639-3: agj

Population	10,860 (1998 census). 44,737 monolinguals. Population includes 47,285 in Amharic, 3,771 in Oromo, 541 in Tigrigna (1998 census). Ethnic population: 62,831 (1998 census).
Region	Fragmented areas along the Rift Valley in settlements like Yimlawo, Gusa, Shonke, Berket, Keramba, Mellajillo, Metehara, Shewa Robit, and surrounding rural villages.
Dialects	Ankober, Shonke. It is reported that the 'purest' Argobba is spoken in Shonke and T'olaha. Lexical similarity 75% to 85% with Amharic.
Classification	Afro-Asiatic, Semitic, South, Ethiopian, South, Transversal, Amharic-Argobba
Language use	3,236 second-language speakers. The ethnic group near Ankober mainly speaks Amharic; the group near Harar mainly speaks Oromo. The ethnic group is working to foster ethnic recognition. Speakers also use Amharic or Oromo.
Language development	Literacy rate in second language: 16.4%.
Comments	Traders; agriculturalists. Muslim.

Entities and Relations => formal data?

- **Relations**: You can define relations (sometimes maybe also difficult, e.g. Gaoshanzu)
- **Entities:** Characterizations, but no definitions
- Systemic or structural entity definition is diacronically seen unsuitable, e.g. Yugoslavia
- 5 Years mapping of language data in XNLRDF.
- Unambiguous data block communication

Entities in Formal Data

- **OLAC**: Don't harmonize data, standardize metadata, don't worry about the meaning of metadata.
- **XNLRDF:** Harmonize, systematize, remove ambiguities
- **Language Monitor:** Don't harmonize or standardize, keep ambiguities, but show source and implictions
  - Traditional facts on entities are theories, ideologies, simplifications, claims
  - Facts in language documentation are documents, not their contents, Meta-data are geo-references or URLs to documents

Avoid Undecidability, Give up Consistency

- Terms have more than one definition, e.g. Yugoslavia, China, English
- A language monitor has to be multi-vocal. If you want people to contribute their data, than you have to accept their 'ontology'
- If you want to distribute data, make clear what theories and their implications are
- Geo-references + time + URL as only common reference system

Avoid Avoid Russel's Paradox, Goedel's Undecidability

- Is each languaging 'Argobba' or is 'Argobba' a property of the class of languagings?
- If each languaging is Argobba, the languaging defines and redefines Argobba, so that the term becomes useless as Metadata.
- Language identifier in a natural language
- Understanding terms: Different hermeneutic traditions, Hirsch (author), Gadamer (subject), Ricoeur (reader).

Data, Meta-data, models and theories

- Data are: documents containing recordings, transcriptsions, maps as unanalyzed entities (not right or wrong)
- Meta-data (time and place, URL)
- Sub-Models about possible relation of types in theories, not a consistent system
- Sentences are instantiations of a sub-models
- Theories are asserted sentences (involving an agent)
- Insights are dynamically constructed from sentences

Examples of Data and Theories

(with special refence to geo-spatial data)

- Photo with Tombstone <=> Time & Space
- Document with Languaging Event <=> Time & Space
- Languaging <=> Time & Space
- Reference in Languaging<=> Time & Space
- Spatial folk taxonomy <=> Spatial Object
- Languaging => Language => Time & Space
- Language => Speakers => Space, Time & Numbers
- Community <=> Time & Space

From Theories to Insights

- Theories should not be merged, insights are dynamic generalizations, conflicting theories, agent-sorted theories.
- Legacy data, are not target models, but serve as base for abudction of sentences, e.g. point-based language mapping).
- Overlap or mutual corrobation calculated: compatible model, compatible vocabulary („many“ <=> „mayority“)

Agents in A Language Monitor: Users and Groups

The I*nformant* provides an insider view.

The A*uthor* is the person publishing his or her thoughts in his or her words.

The Editor provides the information to be published in the language monitor. The editor can be the *author or* cite the author.

The *Target agent* is affected by the information.

The *Reader* is that agent to whom this information should be made available.

Agents in A Language Monitor: Rights and Obligations wrt. models

The *Informant* can be anonymous to readers, can see own contributions

The A*uthor* associated with citation and copyright. Famous Linguist, Normation Institution, Publishing house can be drawn into the system as authors.

The Editor must cite informant or author

The *Target agent* can request to withdraw sentences of some models, probably not all models

The *Reader*. Some agents cannot be excluded as readers.

Conclusion

- Multiple theories, allowing for multiple contributions and reasonable generalization,
- Turn a pseudo consistency into a representation of a discourse, which is as relevant as the content of the discourse
- GPS/GIS root facts, facts root theories
- GPS/GIS can be used to work on theories, but doesn't change their status.