WP2 Lexicons in ECHO
WP2-TR07-2003 - Draft Version 2

Report of the ISO Preparation Workshop on Lexicons

Nicoletta Calzolari, Peter Wittenburg
Febryary 27, 2003

At February 24/25th a workshop on Lexicon Schemas, Registries and Repositories was held in Munich. It was intended to serve as a preparation workshop for the necessary discussions about lexicon related objectives in ISO TC37/SC4. Since text technology is a highly relevant component for all text technology in cultural heritage related work, the ECHO project was stimulating and organizing this workshop. This report will give a summary of the results of this workshop.

Participants of the workshop were:

- Anthony Aristar — Wayne State University Michigan, EMELD
- Nicoletta Calzolari — ILC Pisa (It), ISO TC37/SC4
- Nuria Bel — GILCUB Barcelona
- Hennie Brugman — MPI for Psycholinguistics Nijmegen, ECHO (NL)
- Sebastian Drude — Free Univ of Berlin
- Helen Dry — Eats Michigan University, EMELD
- Scott Farrar — Univ of Arizona Tucson , EMELD
- Brian Fuchs — MPI for History of Science, ECHO Berlin
- Nancy Ide — Vassar College Poughkeepsie (NY), ISO TC37/SC4 (had to cancel)
- Wim Peters — Univ of Sheffield (UK), Ontoweb
- Gregor Thurmayr — ?? (had to cancel)
- Laurent Romary — LORIA Nancy, ISO TC37/SC4
- Gary Simons — SIL International Austin, EMELD
- Christina Vertan — Univ of Hamburg
- Peter Wittenburg — MPI for Psycholinguistics Nijmegen (NL), ECHO, ISO TC37/SC4

The participants came from different areas of applications such as Language Engineering, Field Linguistics, Semantic Web and Cultural Heritage.

The major question raised was: what can or has to be done with respect to lexicons within ISO TC37/SC4 and how can domains such as cultural heritage take profit from the developments towards better standards.

It was agreed that a small group of persons has to be identified soon that takes care of progress in this area of the work within ISO TC37/SC4. The chairman will be asked to make suggestions that also include experts from other continents. Also a broader group of interested and committed experts should be determined that contribute to the work.

It was concluded that amongst the participants there is a lot of agreement with respect to visions and future plans. This commonality should be documented soon to be able to distribute this as a kind of basic position paper. It can be based on Gary Simons' note presented at the EMELD workshop, but has to go beyond.

Pre-conditions
A number of pre-conditions were mentioned that should guide further work:

- The work must have relevance for different application areas such as Language Engineering, Field Linguistics, History of Science, etc. There are overlapping interests but also some differences were identified (flexibility, dynamics, amount of formal definitions, amount of dependencies, ...). It was argued that these differences should become more obvious by practical work to better understand the requirements.
- The work must be relevant for academic and industrial applications.
- Researchers from different areas of the world must be included to also get the requirements from different languages. Here the domain of field linguistics was seen as very interesting, since they cover a large variety of languages.
- The relevant initiatives such as MILE, EAGLES and EMELD (to just mention a few) have to be included and their results have to be understood and used.
- The lexicon may not be seen as an isolated linguistic data type, but in its interaction with texts, grammars, metadata etc.

Relevant Data Types and Terminology
Much time was spent on the question which data types are relevant with respect to lexicons playing a role in the Semantic Web era and what kind of information they should contain.

- It was concluded that the terms that the lexicon experts are using are not at all clearly defined. Therefore it would be very useful to have a kind of glossary with short descriptions of all relevant terms.
- Finally it was agreed that in future it would be preferable to separate the definitions of **linguistic data categories** and its usage for example to establish relations. So a data category repository should contain a list of appropriately defined linguistic concepts where each is described by a number of attributes as suggested by existing ISO proposals/standards. Although it was seen as utterly useful to have different approaches at this moment there were at least two arguments that suggest to separate for example datcat definitions and ontological information: (1) It was assumed that there will exist several ontologies that make use of a given set of datcats. (2) A separation between the different types of information is preferable from the management perspective.
  It was discussed what the difference between primitive and complex datcats are and how they can be described. Complex datcats such as "semantics" that may stay for a sub-block of lexical information cannot be defined by a conceptual model, although people agree on abstract definitions. Due to their unspecificity people can associate different set of lexical attributes and different structures with them. This cannot be part of the definition.
  It was agreed that much clarification is necessary.
- The role of **termbases** for lexicographic concepts was not very clear from several reasons. In general sense it was assumed that termbases will contain associations of datcats to different languages. However, it is clear that many terms used to denote a linguistic concept in various languages may be misleading, since the definition may be different: As example "gender" was mentioned. The concept "gender" stands for an abstract concept the exact definition of which is different in various languages.
  Nevertheless, it is important to support the notion of gender despite the language differences. It was argued that there should be a framework that allows specifying these differences, but on the other hand preserving them. The question was raised in how far differences could be described by associating different features.
- In contrary an **ontology** (referring to linguistic concepts) should include relations of different sort that exist between lexicon datcats. It was agreed that RDF/OWL are the standards that have to be used for defining the relations. Datcats are just seen as points of reference that indicate identity to the inference machine. It was assumed that there will be many "practical" ontologies that may

define relations between datcats and that also various datcat repositories will exist. They can nicely be differentiated by the namespace mechanism.

- Gary Simon proposed the **metaschema** concept. This is important since there are many legacy repositories that use terms that are not properly defined. These have to be linked.
- It was agreed that XML, RDF, OWL will be used as description formalisms.
- It was also agreed that we need good examples to be worked out to make clear what the different data types mentioned contain and how they relate to each other.
- Attempts such as within the EMELD project are very helpful since they will work as testbeds for methods. Their experience should be reported.
- It was expected that the ISO chairman will help clarification by stimulating experts to work out short notes with proper descriptions and examples.

Application of the MILE Model

It was agreed that it is important to try out the MILE model to get practical experience. The MILE model will be applied to various existing dictionaries by creating MILE compliant entries. This voluntary work of a few participants should include the definition of the datcats used and where possible the creation of mappings to existing core datcat definitions. It was expected that this work will improve our understanding about the types of mapping relations needed for representing lexical knowledge.
The following institutions expressed interested in participating in this endeavor: ILC Pisa, U Sheffield, Univ Hamburg, DFKI, MPI Nijmegen)

Merge of Data Categories

It was seen as very important to create datcat repositories with a critical mass of definitions to make it useful for linguists to reuse them. Further, it was argued that we need to merge the existing EMELD, MILE and other repositories to understand the complexity we will be faced with.
It would be very useful to also include for example Asian languages in this exercise due to their different nature.
A few names were mentioned, but this has to be refined (EMELD, Pisa, Drude, Enfield(?).

Registration of Lexical Service

It was agreed that we need a suitable environment for registrating the different lexicon offerings by making use of the upcoming registration services. It has to be investigated whether UDDI with its supporting standards WSDL and SOAP are an option. If so it is clear that “linguistic infrastructure” such as a suitable taxonomy or suitable descriptors for searching have to be added. These have to be worked out. Further, recommendations have to be worked out about the granularity of such services and about the specifications of the input and output parameters.
It was agreed as well that we need small test scenarios as soon as possible to understand the details. A few institutions showed interest in this work: MPI Nijmegen, MPI Berlin, U Sheffield.

Distributed Scenario

The future scenario of lexical services will be a distributed and open one, i.e. lexicon elements will be held at different locations. It has to be worked out in more detail, what the goals, requirements and use cases are for such scenario to not end up in unmanageable infrastructures. The interplay between “core” information provided by trusted linguistic institutions and extensions by other interested persons has to be studied as well as the processes of qualification and merging that may enrich the core information by integrating useful proposals.
The semantic web formalisms have to be applied to achieve this and also here we need small test scenarios as soon as possible. Again a few institutions expressed their interest: U Barcelona, U Arizona, U Hamburg, U Sheffield, DFKI.

Tools

It is clear that we need a new set of tools to support the distributed scenario with rich and diverse data types and where different mappings may be used. This issue was seen as too premature to make more detailed statements and to start practical work. Existing and coming projects may function as test cases for improving our understanding of the requirements.

Organizational Aspects
Also this aspect needs deeper analysis. A few aspects were mentioned that have to be tackled to make the kind of open and distributed scenarios working: Quality control, validation, new business models, maintenance, sustainability.
Also the point of the nature of the formal ISO procedure was mentioned as a question to be clarified.

Next Workshop
It was agreed to organize a next workshop around September/October 2003 as a follow up. This workshop should include more people - also from other regions of the world and some test cases and experiences should be reported. So the workshop was seen as a motivation to start the work mentioned above and to report on it.