LREC-2004 WORKSHOP: "A REGISTRY OF LINGUISTIC DATA CATEGORIES WITHIN AN INTEGRATED LANGUAGE RESOURCES REPOSITORY AREA"

DATA CATEGORIES IN LEXICAL MARKUP FRAMEWORK OR HOW TO LIGHTEN A MODEL

Gil FRANCOPOULO AFNOR-INRIA gil.francopoulo@wanadoo.fr
Monte GEORGE ANSI dracalpha@earthlink.net
Mandy PET ANSI-ORACLE mandy.pet@oracle.com

1 Introduction

Previously, ISO TC37 efforts have focused on standards and associated models for dealing with language resources such as terminologies but up until now focus has not been on the various other aspects of language processing. The Lexical Markup Framework (LMF), a proposed standard numbered ISO-24613, addresses lexical resources at a higher level that allows for interoperability with terminological, human oriented lexical resources and machine-based NLP lexica. LMF relies heavily on the ISO-12620 data categories (DC), managed under the auspices of the ISO-12620 revision by Laurent Romary (AFNOR-INRIA). They serve as lego building blocks used to facilitate this operability.

We will see how the DC ease the definition and use of various norms and particularly lexical models.

2 Current situation

Traditionally, concerning linguistics constants, the two following strategies are applied:

Strategy #1:
The lexical model defines the list of all the possible values for a certain type of information. For instance, /gender/ could be /masculine/, /feminine/ or /neutral/.
More precisely, there are two sub-strategies:

- define that /gender/ is /masculine/, /feminine/ or /neutral/ without any more details.
- define that /gender/ is /masculine/ or /feminine/ for French and /masculine/, /feminine/ or /neutral/ for German.

Strategy #2:
The values are not listed at all. The model just states that there is the notion of gender.

An example of the first strategy is applied in the GENELEX [Antony-Lay] and EAGLES models where the DTD contains all the possible values. The drawback of such an approach is that the DTD is necessary huge and could be incomplete, specially for languages unknown to the model authors.

The advantage of the second strategy is that the model is simple and nothing is forgotten. But its drawback is that such a model is useless and we will see that in the next paragraph.

3 Capacities

For a lexical model, we can distinguish two criteria:

- The power of representation: what kind of data the model is able to represent ? what language the model could be applied to ?

- The power of operation: is it possible to compare two words ? how to present a pick list to a user of an interactive workstation ? is it possible to merge two LMF conforming lexica ?

The two criteria are somehow contradictory: the more generic the approach, the more diverse lexica are needed to merge.

Coming back to the second strategy that is to avoid defining the possible values for gender, the power of representation is high but the power of operation is very low. Nothing guarantees that a lexicon defines gender as /m/ and /f/, or /mas/ and /fem/ or worth /neuter/ for French. **In such a situation, comparing words or merging various lexica are difficult operations and the norm becomes useless**.

4 Merging

Let's detail a bit what is merging.

Merging can take various forms such as the following use cases:

Use Case #1
Situation: Multilingual lexicon in N languages
Goal: Add 1 new language to this lexicon

Use Case #2
Situation: Monolingual lexicon in language L
Goal: Add words in language L

Use Case #3
Situation: Multilingual lexicon in N languages
Goal: Add missing translations

Let's add that merging is a frequent operation and is an heavy burden for the lexicon manager.

5 Solution

The solution is not easy. We must represent existing data and due to the extension of multilingual databases and various formats used, merging seems to be the most demanding operation.

There is another point to be mentioned. This problem is not specific to lexicon management. The gender definition is shared by other processes like text annotation and features structures.

That means that:

- It is not very wise to duplicate the effort in various norms.
- Text annotation, features structure coding and lexical representation are not independent processes. In case of parsing for instance, the information extracted from the lexicon will be transferred to annotation or feature structures, there is the danger to produce different (and so incompatible) values.

The solution is to define data categories in a separate norm. These values will then be shared by the lexicon, annotation and features structures norms. And of course other future norms could take place in this architecture.

6 Details

The data categories are not only constants like /masculine/ preferred to /m/ or /mas/ but are defined according to the language processed.

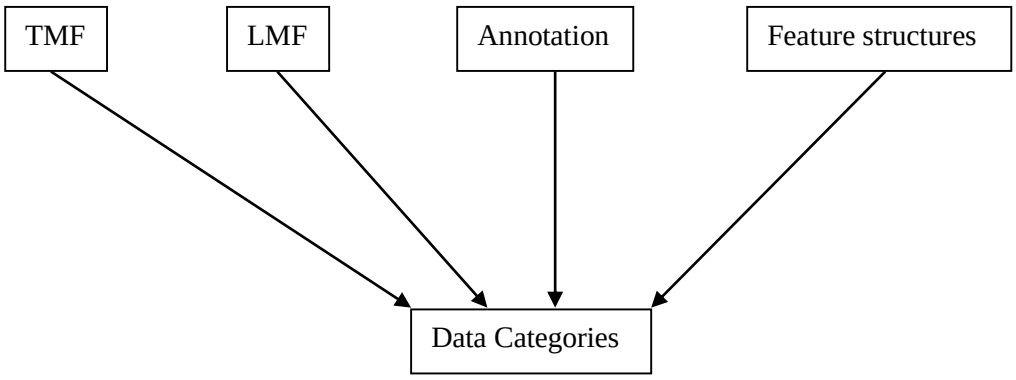
More precisely each feature will be defined as a tree. The top node is /gender/ for instance. One level below, we have /french/ and the possible values are /masculine/ and /feminine/. At the same level as /french/, we have /german/ and the possible values are /masculine/, /feminine/ and /neuter/.
For an unknown language, the possible values are the union of all values extracted from all languages.

As it could be noticed, the number of values is quite important. A management tool is needed in order to ease data category search and selection. Such a tool is provided by INRIA under the auspices of the Syntax project.

7 A family of norms

The process used is similar to the one of TMF (aka Terminological Markup Framework) that is the ISO norm for thesaurus [Romary].

Data categories are located at the lower level of the TC37 family of norms as sketched in the following diagram.



And the four norms are based on data categories, so each norm is light, non redundant and can interoperate with the others.

8 Conclusion

Like the other norms of the family, the base line for LMF is to:

- Concentrate on structuring the elements and linking elements together.
- Relegate language idiosyncrasies in an external and shared norm: ISO-12620.

As we have seen, LMF is part of a more global ISO move in order to define a set of coherent norms based on data categories.

Bibliography:

Antoni-Lay M-H., Francopoulo G. and Zaysser L. 1994
A generic model for reusable lexicons: The GENELEX project.
Literary and Linguistic Computing 9(1): 47-54.
Romary L. 2001
Towards an Abstract Representation of Terminological Data Collections – the TMF model.
TAMA. Antwerp.