Minutes of the LIRICS lexicon meeting
ILC-CNR Pisa 16th February 2005
Version 2 of the document

Attendees:

Monica Monachini	CNR
Nuria Bel	UPF
Nicoletta Calzolari	CNR
Claudia Soria	CNR
Mark Kemps-snijders	MPI
Gil Francopoulo	INRIA

Author of the minutes: Gil Francopoulo with review by attendees

Scheduled agenda:

- syntax
- predicates or not predicates
- derivation
- phonology
- unified lexicon (connection with LC-Star)

Minutes:

1) Organisation of the LMF document

We decided to have the following structure:

a) Core model

b) 2 Extensions MRD and NLP

Each group of classes will be explained by the mean of a class UML diagram. The diagram is followed by an extract of an example.

c) Full examples

The examples are presented from the most simple to the most complex one. Each examples are presented with a text, a UML instance diagram (in order to easily correspond to chapters a and b) and the XML data. We don not know in what XML style the data will be presented. The first example is an invented one and the others are real entries of existing lexicons.

- Getting started example
- Italian Parole lexicon
- WordNet
- LC-Star
- FrameNet
- others: ask Monte if he wants to add MRD real entries

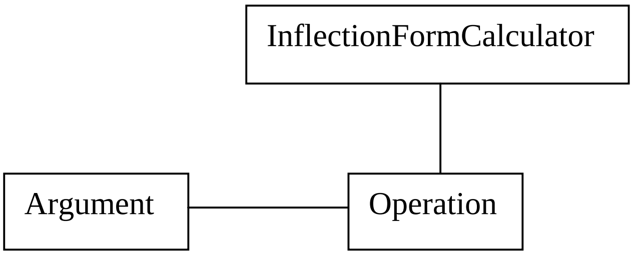
2) General

We all agree that the LMF document must be driven by linguistics concerns: how to represent such and such linguistic phenomena. The meta model and the software tool must follow, and not the contrary.

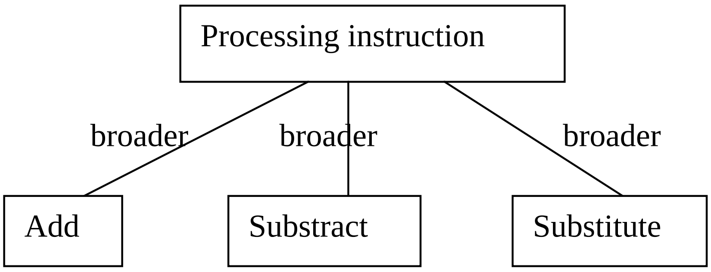
3) NLP Morphology

Instead of having a fixed list of operations, we decided to have the notion of operation and argument. It's more generic and more concise to explain. The operation will be defined by a data category (DC) called /processing instruction/ that has narrowed DCs like /add/, /substract/, /substitute/ and so on. The class operation will have a content that is a DC of type /processing instruction/. The Argument class will have a content that is a string. The association between InflectedFormCalculator and Operation is ordered. The association between Operation and Argument is ordered.

Something like this in the model:



And something like this in the DCR:



4) NLP Syntax

Concerning the relation between syntax and semantic, the situation in LMF and Eagles has certain drawbacks. In Eagles, the syntax is mandatory: we cannot describe a sense without any SyntacticBehavior. In the actual LMF, it is the contrary: we must have a semantic unit before having a SyntacticBehavior.

Nuria said that certain actors use a shallow syntactic description to record patterns used after for named entity recognition (NER). And NER is very popular in Text mining and Information extraction in general. Such actors do not have necessarily very huge semantic layers.

Nicoletta said the problem of having to fix the semantic decomposition in different units is a difficult problem, because for a given word, at the start of the word description, we do not know how many units we will have: on the contrary, we know the syntax the word. So fixing the syntactic behavior after the semantic decomposition means that some book keeping and reorganisation will be mandatory after a more precise semantic study.

Gil says that the situation in Eagles where syntax is mandatory produce other problems. A lot of actors have morphology and semantic without any syntax. On the contrary, and specially in the private sector, the actors that have a syntactic layer without any semantic are a very small minority: if they exist at all. Even the actors in the MT field that operate on a transfer based approach (i.e. the translation connection is made at the syntactic level) have at least very simple semantic features like +human etc.

So we decided to permit the direct connection between LexicalEntry and SyntacticBehavior.

We must find a way to connect SyntacticBehavior to Sense: may be an object called a connector could do the job.

5) NLP Semantic

At the moment, the Relation regroups the notion of relation (like in Eagles, a semantic relation) and a feature. A feature is just defined as a relation that has no target. We agree that this is not very explicit. So, it's better to have (like in Eagles) the notion of relation and the notion of feature.

We decided to have predicates. We did not decided of a complete and precise model concerning the predicate. The question is : do we use 'as is' the Eagles model that is quite complex, or do we elaborate something a little bit simpler. The answer is: study the problem within the next few days among us.

It has been noticed that the two classes: definitionBlock and Proposition could possibly be factorized with the predicate. We have to be sure that the new model will be able to represent the DEC.

Nicoletta noticed that we must take care of being able to represent FrameNet.

Monica and Gil agreed to say that is quite easy to represent ComLex.

6) NLP Derivation

Nuria said that she uses the predicates for describing derivation relations.

Nicoletta and Gil agreed to say that representing derivation in morphology instead of semantic goes to a dead-end at least for European languages. Derivation is proposed in the Eagles morphology but as far as we know nobody uses it. Everybody in the Genelex, Eagles, Parole, Simple network describes derivation in semantic.

Gil said that the only interest for derivation in morphology is for the ones that do not have any semantic information. It could be interesting for Arabic language where derivation is very productive. The explicit representation of derivation seems to be a valuable mechanism for learners of Arabic language.

7) Arabic language

We must provide a mean to record a root in the lexical entry for Arabic language. There are (at least) three strategies concerning the representation of morphology for Arabic language. The problem is to manage the voyelled form from the root.

In Egypt, E* starts from the root and uses a statistical matrix to add voyelles. This gives very small programs but with some noise.

In the US and France, X* a couple of years ago tried to use transducers (with may be the most talented experts in transducers) in order to compute the forms from the root and it's known to be a failure.

Recently, in France, F* now seems to succeed in considering the root as a comment that is only useful for human beings, but the voyelled form is recorded in order to be the source of the computing. This actor learnt from the past experiences and of course this takes a lot of space but it works. The voyelled form is used as a lemmatised form like in the other languages. This is not the way human beings learn the language but the goal of this actor is to build NLP systems that work, nothing else.

8) NLP Interlingua

What is missing is the possibility to represent something like: “transfer only valid for medicine”. In other terms, we need the possibility to express constraints.

The transfer axis should have an attribute that permits to specify that this transfer is bidirectional or unidirectional.

Other items:

- We all agree that we speak the “same language”. It’s not only because we have the same background but also because we have the same concern: “represent words, their behaviors, their meanings and translations for NLP”.
- We must study the project of declaring the pronunciation within W3C. Nicoletta will send the message announcing this work. This project bases its work on ISLE, OWL and the RDF file written by Nancy.
- Within the W3C people, WordNet is considered (oddly) as an ontology.
- Monica and Nuria ask for a training on the Syntax tool that is not intuitively usable. Two hours of training by Philippe should be fine. Gil said that he does not know all the details of the tools.