The relevance of standards for research infrastructures

Gil Francopoulo[1], Thierry Declerck[3],

Monica Monachini[2], Laurent Romary[4]

[1]INRIA-Loria: gil.francopoulo@wanadoo.fr
[2]CNR-ILC: monica.monachini@ilc.cnr.it
[3]DFKI: declerck@dfki.de
[4]INRIA-Loria: Laurent.Romary@loria.fr

Abstract

In this paper, we show the importance of standards as an essential aspect for any research infrastructure in the humanities. In the context of the current activities within ISO committee TC 37/SC 4 (Language Resource Management), we show in particular how important it is to provide means to compare linguistic representations through the use of a shared semantics for elementary descriptors. This is further exemplified by describing the ongoing work to define a central *data category registry*, which aims at being a reference point in the language resource community, in conjunction to the definition of basic standards for linguistic annotation, as illustrated with the current work that is being carried out in the domain of morpho-syntactic categories.

1. Standards: are they at all needed ?

For many years, the language resource community has been the place of numerous projects (see Cole et alii, 1997) that have aimed to produce resources and tools to facilitate the study or automatic processing of language. Still, we have all faced the issue of ensuring long-term availability of the corresponding results, with the consequence that researchers still have to carry out technical tasks of corpus gathering, lexical description or tool implementation that others are supposed to have achieved beforehand, and above all that should be the duty of shared research infrastructures working for the benefit of all.

One of the key issues to define such research infrastructures is our ability, as a mature scientific community, to be able to identify that new research results should be based upon the stabilization of shared knowledge by means of a range of internationally agreed upon standards. Such standards would obviously bring the following benefits:

- Ensure wide accessibility of data in space (between research sites) and time (in the perspective of providing long-term preservation of data). Standards are there to provide a stable representational basis as well as maintained documentation, that researchers are not able to produce on their own;
- Facilitate the reusability of software by making it independent from the actual proprietary data formats an implementer might use;
- Guaranty that research results are comparable, by, for instance, making sure that the same underlying data has been used in the context of the elicitation of statistical results;
- Create communities of practice that will share the knowledge of such standards and create new concepts on the basis of this common culture.

As a matter of fact such benefits have already been observed in the context of the wide deployment of the Text Encoding Initiative guidelines, which have both been the basis of numerous projects worldwide[1], but also have been the basis of a shared understanding of basic textual descriptions that now leads to the explorations of new textual types or phenomena[2].

Still, the language resource community requires even more standards to cope with both the variety of linguistic phenomena that have to be taken into account as well as the diversity of human languages. This is why, a the International Organization for Standardization[3] has put together a new committee dedicated to language resources, known as ISO/TC 37/SC 4 and started to foster several standardization projects to deal with what has been identified as priorities for the progress of the management of language resources.

In the remaining sections, we first provide a few elements related to the role we think research infrastructures should play with regards standards. We then outline the working agenda of ISO/TC 37/SC 4 and we present our opinion concerning standards when applied to Research Infrastructure (RI). Then, as an illustration, we present the work in progress within ISO-TC37/SC4 on the morpho-syntactic profile of the data category registry (DCR).

2. Research infrastructures and standards

As we have seen, standards are an essential component of any language resource related activity. In this context research infrastructures should consider standardization as one essential point of their activities. More precisely we consider that at least the three following missions should be allocated to research infrastructures:

- They should contribute the wide dissemination of standards by initiating training sessions and providing teaching materials and samples on line;
- They should actually implement available standards in all their activities, with the constant objective of

[1] See the TEI projects page under http://www.tei-c.org/Applications/

[2] See the P5 edition of the guidelines: http://www.tei-c.org/P5/

[3] http://www.iso.org

long-term availability of the data or tools they produce (see above);

- They should be at the forefront of standardization activities by explicitly reviewing existing standards, contribute to their evolution and even participate to the definition of new standards when needed by the corresponding research community.

3. Work in progress within ISO-TC37

ISO committee TC 37/SC 4 is dedicated to the specification of a full family of standards for NLP and language resources. These standards can be categorized according to two levels:

Low level standards, describing the linguistic constants. More precisely, this is a pair:

a) revision of ISO-12620 that specifies the rules for describing and maintaining data categories.

b) data category registry

There are also some other important low-level standards that we can use: the standards for character encoding (ISO/IEC 10646 i.e. Unicode), language codes (ISO-639), script codes (ISO-15924), country codes (ISO-3166) and dates (ISO-8601).

High level standards, describing structural models (sometimes called meta-models) that specify how to represent linguistic resources. The structural model provides classes (in UML terminology) and the relations between classes together with a textual usage description for each class.

The registry provides the needed attributes and values that are used **to adorn the classes**. The structural models being currently developed deal with word-segmentation, morpho-syntactic annotation (aka MAF), syntactic annotation (aka SynAF) [1] and lexicon (aka LMF) [2].

4. Objective

The objective is to propose to the user and developer of language resources a coherent family of standards. All these standards have the following property: they allow the definition of a model of linguistic resource by combining structural elements with constants taken in low-level standards. All the resources share thus the same set of constants, supporting our goal of providing interoperability between segmentation, annotation and lexicon.

5. Roadmap

As said before, the duration for defining an ISO standard is rather long. It takes around four years. So, instead of defining low-level standards then high level standards (or the contrary), the various ISO groups works in parallel with a closed collaboration between them.

6. Some basic definitions

6.1. A data category

A data category is a linguistic constant. A data category is either an attribute name like /partOfSpeech/ or a value dedicated to populate an attribute. An example of value is /noun/.

6.2. Profiles

A profile is a specific set of data categories in the DCR.

The current profiles are:

For Terminology within TC37/SC3
One profile
For NLP within TC37/SC4
Three profiles:
Meta-data
Morpho-syntax
Semantics

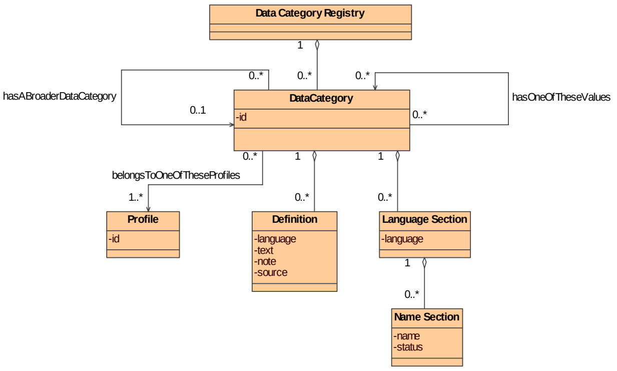
You can notice that to ensure interoperability in NLP between word-segmentation, annotation and lexicon, the distinction between each profile is made according to linguistic criteria and not according to the resources. Another point to mention, is that a data category may belong to several profiles but we try to avoid this situation in order to avoid conflicts.

6.3. The data category registry

The registry is the union of all data categories.

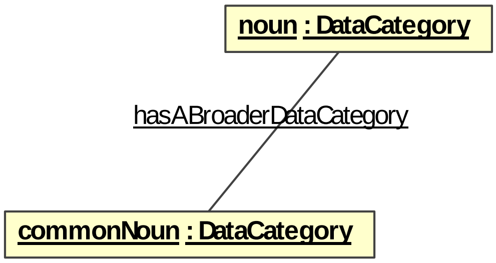
7. Morpho-syntactic profile

The DCR structure is specified by the ISO-12620 revision. In the morpho-syntactic profile we restrict ourselves for the time being to the following features:

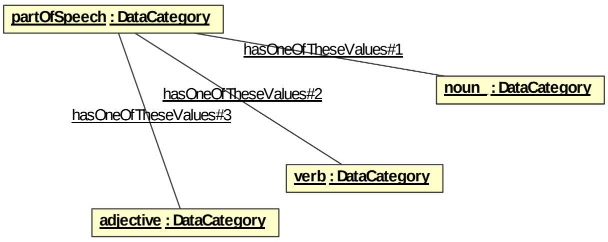


We differentiate between the notion of /broader/ relation and the notion of /conceptual domain/.

The /broader/ link allows a hierarchy of constants to be defined. Example: a common noun is a more specialized value than noun.



The notion of conceptual domain allows a set of valid values to be identified. Example: noun is a value for partOfSpeech.



We proceeded in three phases:

Phase-1: collect

Phase-2: group, structure and write a first draft of the definitions

Phase-3: revise

An initial long and flat list of data categories has been collected from:

- Current ISO-12620
- Eagles and Multext-East

8. What has been done in the morpho-syntactic profile?

- A couple of values for the NLP sections in LMF

The ISO-12620 constants are general purpose values like /language/ or /derivation/ and cover only terminological resources. For instance, for /partOfSpeech/, the only values are /noun/, /adjective/ and /verb/. By comparison, in NLP, we need much more values including /preposition/ and /pronoun/ etc.

We propose a set of constants according to the following criteria:

- broad linguistic coverage within the morpho-syntactic perimeter
- no semantic overlap
- good choice of a name associated with a good textual definition

9. What has been recorded so far in the DCR?

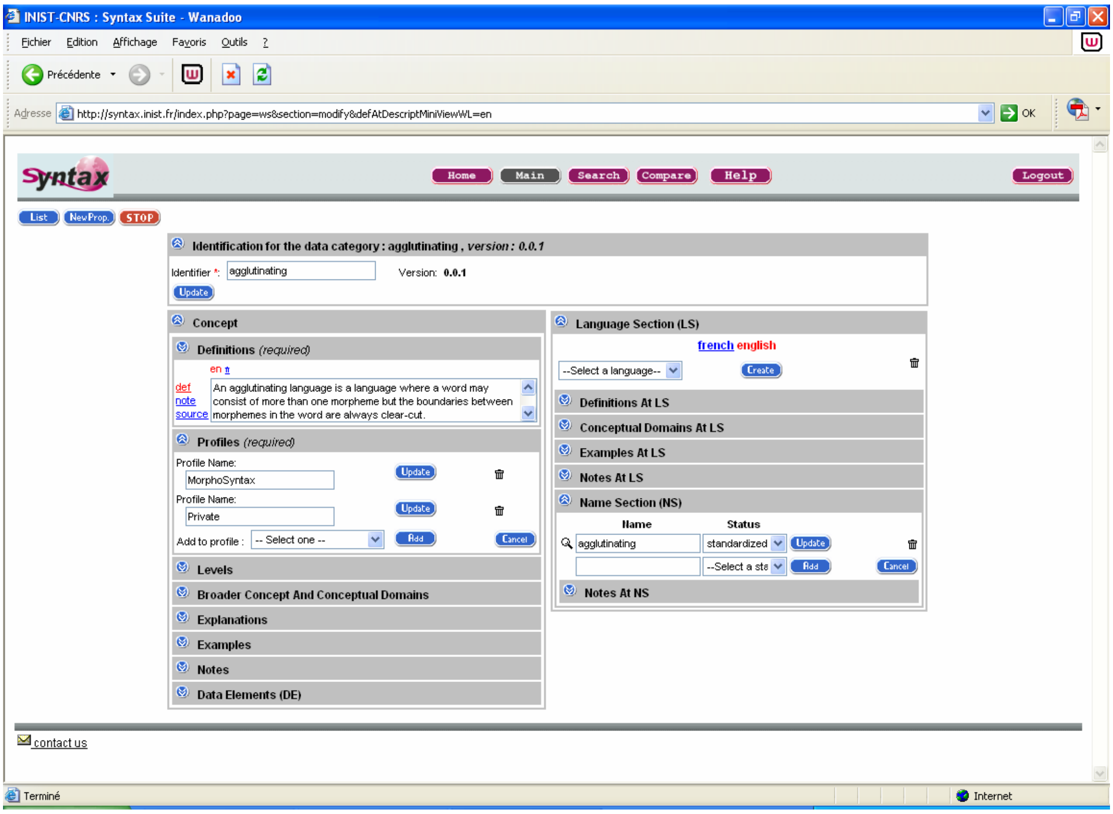
The list being rather huge we created 11 directories within the Syntax software (see next section) in order to help data category organization. It easier to work on medium sized list than on a list with 300 items.

In each directory: one or several attributes names and related values are recorded.

Basics		29	items
	These are general purpose linguistic constants, like: comment, derivation, elision, foreignText, label.		
Cases		33	
	Examples of values: ablativeCase or dativeCase.		
FormRelated		33	
	These are constantes for the specifications of forms like: spokenForm, writtenForm, abbreviation, expansionVariation, transliteration, romanization, transcription, script.		
Language Typology		4	
	An attribute is languageTypology and values are agglutinating, inflectional and isolating.		
Morphological Features excluding cases		72	
	Attributes are for instance grammaticalGender, mood and tense. Values are for instance feminine, indicative, present.		
Operations		8	
	The constants are for instance addAfter, addBefore, copy etc.		
Part of speech		93	
	The part of speech values are structured with a top level set composed of 10 values like noun or verb. A very precise ontology is specified for grammatical words. Most of parts of speech are common to lexicons and annotations but two set of values (i.e. punctuation and residual) are specific to annotation and are not usually used in lexical descriptions.		
Reference		5	
	The constants are anaphora, antecedent, cataphora, coreference, endophora and referent. This is some doubt to maintain these constants in the morpho-syntactic profile.		
Register, dating and frequency		19	
	The constants are slangRegister or rarelyUsed.		
Semantically motivated		16	
	The constants are agent, intensive. This is some doubt to maintain these constants in the morpho-syntactic profile.		
Syntactically motivated		36	
	Attributes are function or voice. Values are subject, activeVoice for instance.		
Total		348	items

10. Software

We use the Syntax software hosted by CNRS-INIST in Nancy (see http://syntax.inist.fr) in order to edit the data categories. This is a server based on a relational database with a set of PHP programs in order to manage the interaction. Here is a screen dump:



http://syntax.inist.fr/index.php?page=ws&section=modify&defAtDescriptMiniViewWL=en

Identification for the data category : agglutinating , *version : 0.0.1*

Identifier *: agglutinating Version: 0.0.1

An agglutinating language is a language where a word may consist of more than one morpheme but the boundaries between morphemes in the word are always clear-cut.

Profile Name: MorphoSyntax

Profile Name: Private

Add to profile : -- Select one --

Name	Status
agglutinating	standardized

11. APIs

In order to allow programs to access to the DCR, a set of Application Programming Interfaces are being specified and implemented by Max Planck Institute for Psycholinguistics of Nijmegen, INRIA-Loria and University of Sheffield.

12. Acknowledgements

The work presented here is partially funded by the EU eContent-22236 LIRICS project[4], and by the French TECHNOLANGUE program[5].

13. References

Cole, R., Mariani, J., Uszkoreit, H., Zaenen, A. and Zue, V. (Eds.) 1997. *Survey of the State of the Art in Human Language Technology*, First Edition – 1997, Cambridge University Press.

Declerck T. 2006 SynAF: towards a standard for syntactic annotation LREC Genoa.

Francopoulo G., George M., Calzolari N., Monachini M., Bel N., Pet M., Soria C. 2006 Lexical Markup Framework (LMF). LREC Genoa.

[4] http://lirics.loria.fr

[5] http://www.technolangue.net