Morpho-syntactic profile

in the ISO-TC37/SC4 Data Category Registry

Gil Francopoulo (INRIA-Loria)

Monica Monachini (CNR-ILC)

Thierry Declerck (DFKI)

Laurent Romary (INRIA-Loria+CNRS)

1 Context

The ISO-TC37/SC4 committee 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:

  1) revision of ISO-12620 that specifies what is a data category (aka datcat), how the datcats are described and maintained.

  2) the registry of datcats (aka DCR)

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

- **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 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.

3 Some basic definitions

3.1 A datcat

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

[1] SynAF has been submitted by a member of the LIRICS project as a new work item.

3.2 A profile

A profile is a specific set of datcats in the DCR.

The current profiles are:

- For TC37/SC3: terminology
- For TC37/SC4: NLP
    - Meta-data
    - Morpho-syntax
    - Semantics

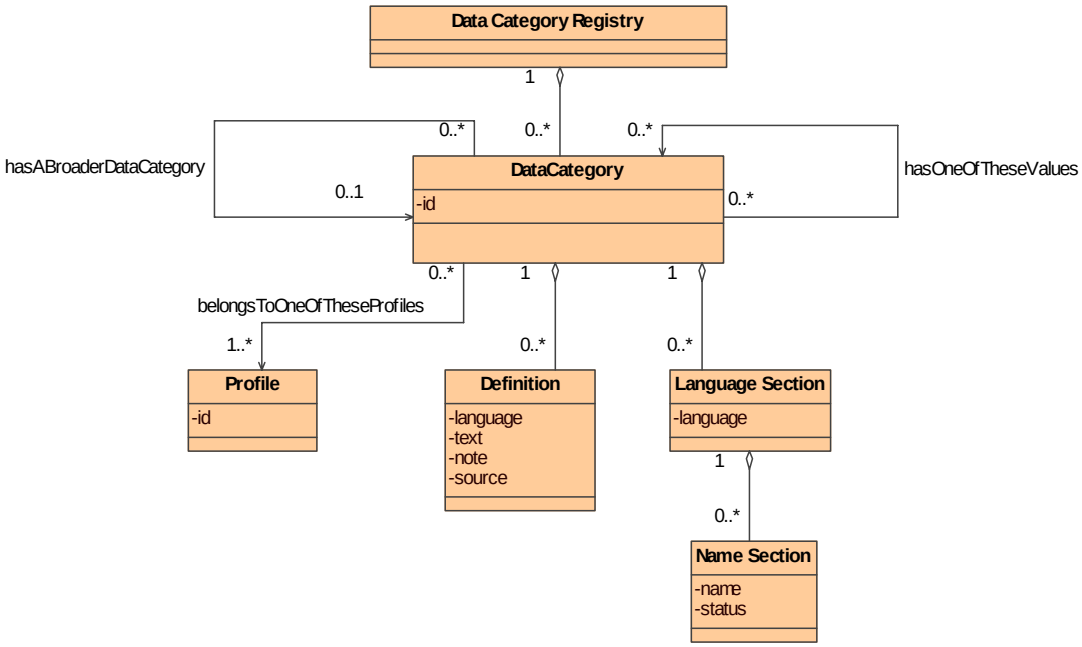
You can notice that to ensure interoperability in NLP between word-segmentation, annotation and lexicon, the distinction between each profile is made according to linguistic criteria and not according to the resources. Another point to mention, is that a datcat may belong to several profiles but we try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

3.3 The data category registry

The registry is the union of all datcats.

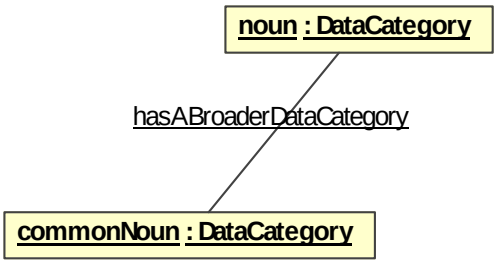
4 The morpho-syntactic profile

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:

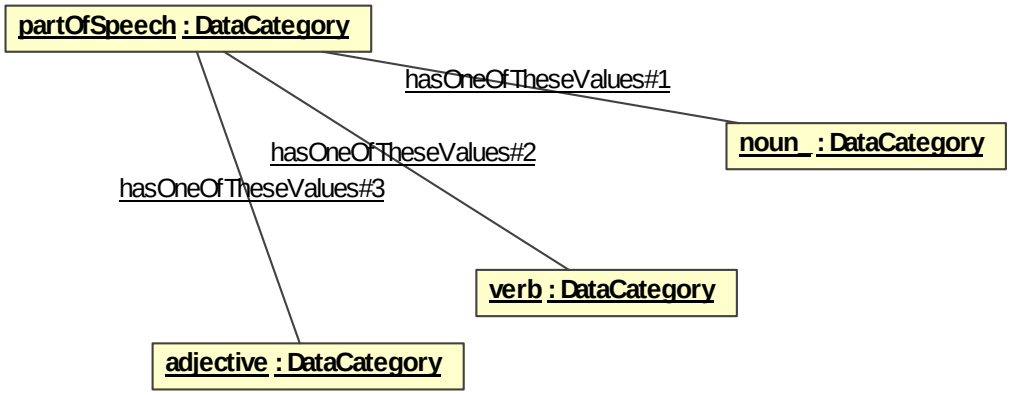


We differentiate between the notion of /broader/ relation and the notion of /conceptual domain/.

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.



5 What has been done in the morpho-syntactic profile?

We proceeded in three phases:

Phase-1: collect

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

Phase-3: revise (to be done)

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

- Current ISO-12620
- Eagles and Multext-East
- A couple of values for LMF

The ISO-12620 constants are general purpose values like "language" or "derivation" and cover only terminological resources. For instance, for /part of speech/, the only values 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

6 What has been recorder so far in the DCR ?

6.1 Directories

The list being rather huge (302 items) we created 12 directories within the Syntax software (see chapter-7) in order to help datacat organization. In each directory: one or several attributes names and related values are recorded.

Basics	40	items
Cases	35	
Language Typology	4	
Morpheme Stem Affix	6	
Morphological Features excluding cases	39	
Operations	8	
Part of speech	84	
Reference	6	
Semantically motivated	19	
Syntactically motivated	38	
Transliteration Transcription	7	
Don'tKnow	16	
Total	302	items

6.2 Basics directory

These are general purpose linguistic constants, like: abbreviation, comment, derivation, elision, expansionVariation, foreignText, homograph, label, native, spokenForm, writtenForm.

6.3 Cases directory

Examples of values: ablativeCase or dativeCase.

6.4 Language Typology directory

An attribute is languageTypology and values are agglutinating, inflectional and isolating.

6.5 Morpheme Stem Affix directory

The constants are affix, infix, morpheme, prefix, stem and suffix.

6.6 Morphological features (excluding cases) directory

Attributes are for instance grammaticalGender, mood and tense. Values are for instance feminine, indicative, present.

6.7 Operations directory

The constants are for instance addAfter, addBefore, copy etc.

6.8 Part of speech directory

The part of speech values are structured with a top level set composed of a dozen of values like noun or verb. A very precise ontology is specified for grammatical words.

6.9 Reference directory

The constants are anaphora, antecedent, cataphora, coreference, endophora and referent. This is some doubt to maintain these constants in the morpho-syntactic profile.

6.10 Semantically motivated directory

The constants are agent, intensive. This is some doubt to maintain these constants in the morpho-syntactic profile.

6.11 Syntactically motivated directory

Attributes are function or voice. Values are subject, activeVoice for instance.

6.12 Transliteration Transcription directory

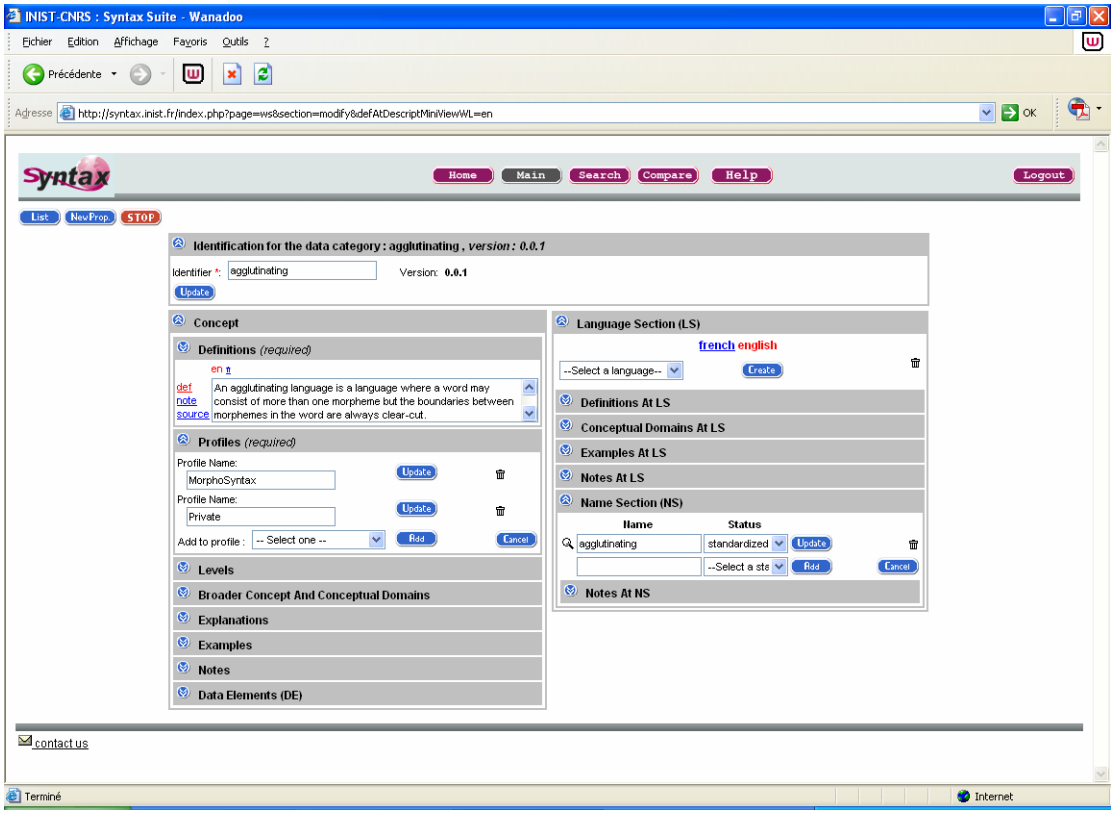
The constants are transliteration, romanization, transcription or script.

6.13 Don't know

This directory lists some constants that are not yet classified like instructive or unique.

7 Software

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



The current software is fine for editing a single datcat but it lacks more sophisticated features related to a set of datcats. Hopefully, the server has an XML export mechanism and so we

were able to implement additional programs[2] for instance in order to list hierarchically the identifiers according to the broader link. Another program verifies that a certain directory complies to certain criteria, so we are able to compute an automatic diagnosis for a given directory.

8 Aknowledgment

The work presented here is partially funded by the EU eContent-22236 LIRICS project. The purpose of LIRICS is to develop and promote ISO standards for NLP language resources.

The work is partially funded by the French TECHNOLANGUE program.

9 Conclusion

The organization in small directories seems fine. In fact, there is no other possiblity: the list being too huge to allow any serious work.

Now every definition must be checked: some are rather fuzzy or incorrect. We also have to check whether the values cover correctly word-segmentation, annotation and non-European languages.

[2] Implemented in Java with the Java SAX parser