Ad hoc Thematic domain group 2: Morpho-syntax

ISO-TC37 meeting Jeju 21 Jan 2006

Gil FRANCOPOULO

INRIA-Loria (LIRICS project)

Summary

- **1 What is a datcat?**
- **2 What is a profile?**
- **3 What is the situation in the morpho-syntactic thematic domain group?**

  3.1 what has been done?
  3.2 what is left to do?

1 What is a datcat?

=> A datcat is a constant

- Context:
- In TC37/SC3+SC4, we have and we are on the way to specify two sorts of standards
- **Low level standards**

- This is the pair:
- Revision of ISO12620 that specifies how the datcats are described and maintained
- And the registry of datcats (DCR)

- This registry will provide all the linguistic constants that we need

- There are also some other important low level standards that we need, but we are not going to define them because they already exist. So we will use them. These are for character codes, language codes,script codes country codes.

High level standards

- These are structural models (sometimes called meta-models) that speficy how to represent linguistic resources.
- The structural model provides the classes (in UML terminology) and the registry provides the attributes and values. These latest are used to adorn the classes.
- These structural elements deal mainly with: word-segmentation, morpho-syntactic annotation (MAF), syntactic annotation (SynAF) and lexicon (LMF).

Objectives: the goal is to propose to the user a coherent family of standards

- All these standards share this property: the user can define a model of linguistic resource by combining structural elements with constants taken in the datcat registry (DCR).
- So all these resources share the same set of constants. The goal is to provide a good interoperability between segmentation, annotation and lexicon.

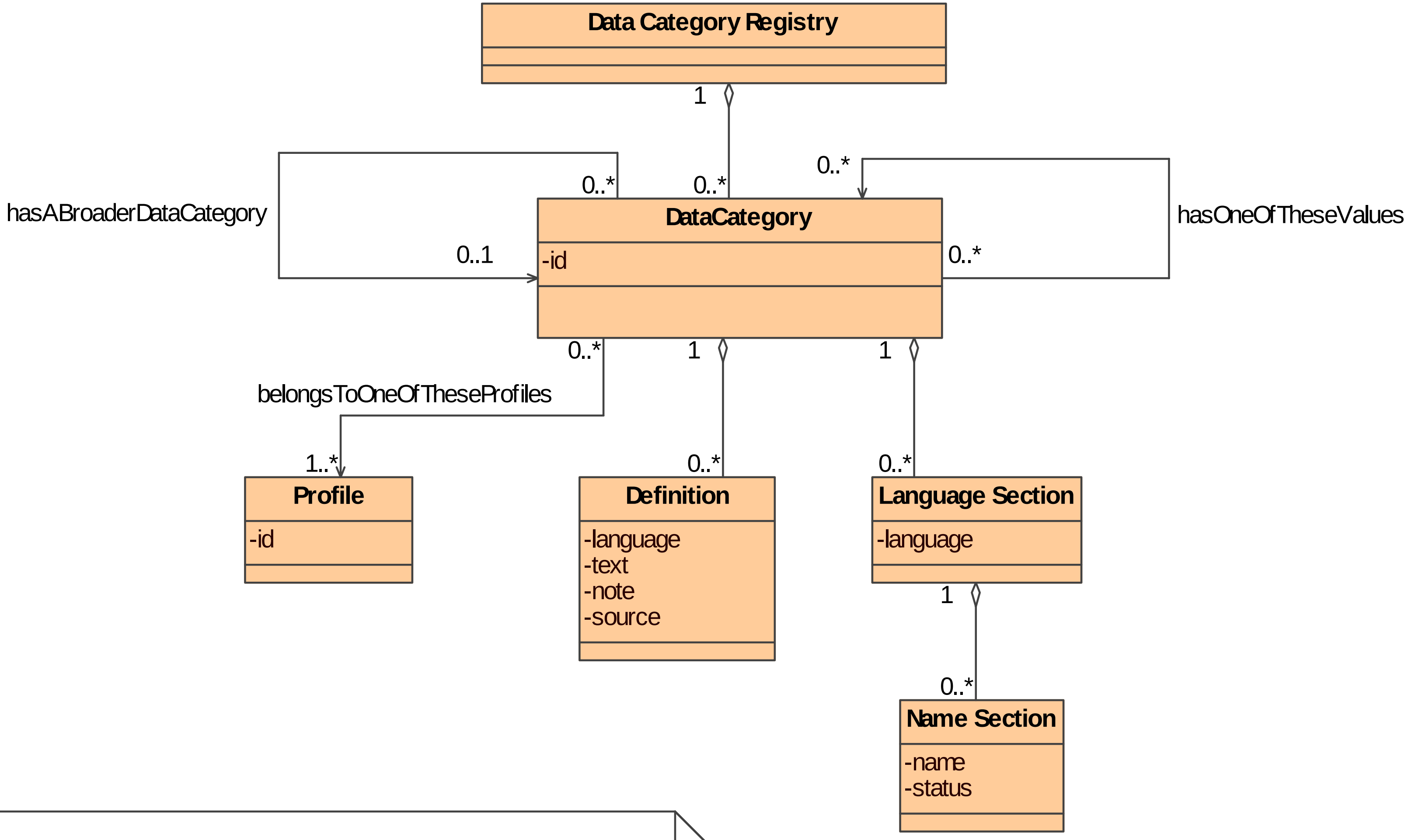
2 What is a profile?

- A profile is a set of datcats in the DCR
- The current profiles are:
  - for SC3: Terminology (for TMF)
  - for SC4: NLP
    TDG1 meta-data
    TDG2 morpho-syntax
    TDG3 semantics
- Note-1: in order to ensure a good interoperability between WS, Annot & Lexicon, the profiles are the same. The split is made on the linguistic criteria, not the resources
- Note-2: a datcat may belong to several profiles but in fact we try to avoid this (in order to avoid conflicts).

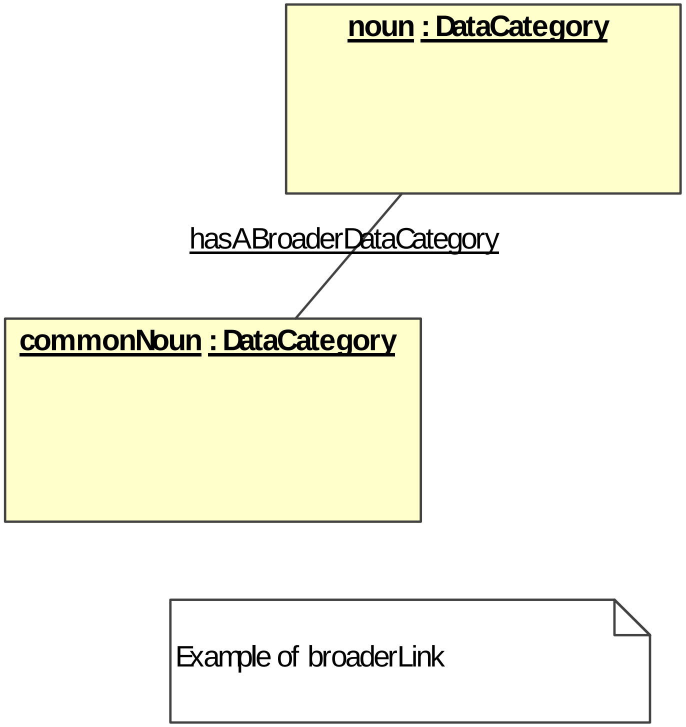
3.1) What has been done in the morpho-syntactic thematic domain group ?

- Progress has been rather slow: 3 phases
- PHASE-1: to collect (done)
- PHASE-2: to group, to structure and complete the definitions (done)
- PHASE-3: to make a revision (to be done)
- PHASE-1: an initiale flat list of 281 datcats has been collected from:
  - current ISO-12620
  - Eagles
  - Multext-East
  - a couple of values for LMF

- The constants coming from ISO-12620 are general purpose values like « language » or « derivation ». But it’s not enough for NLP resources because they cover just terminological resources. For instance, for /part of speech/ the only values are /noun/, /adjective/ and /verb/. By comparison, in NLP we need all linguistic values like /preposition/ and /pronoun/.
- In fact, most linguistic values come from Eagles. And extension for slavic languages comes from Multext-East.



classes and attributes that are used in the morpho-syntactic profile

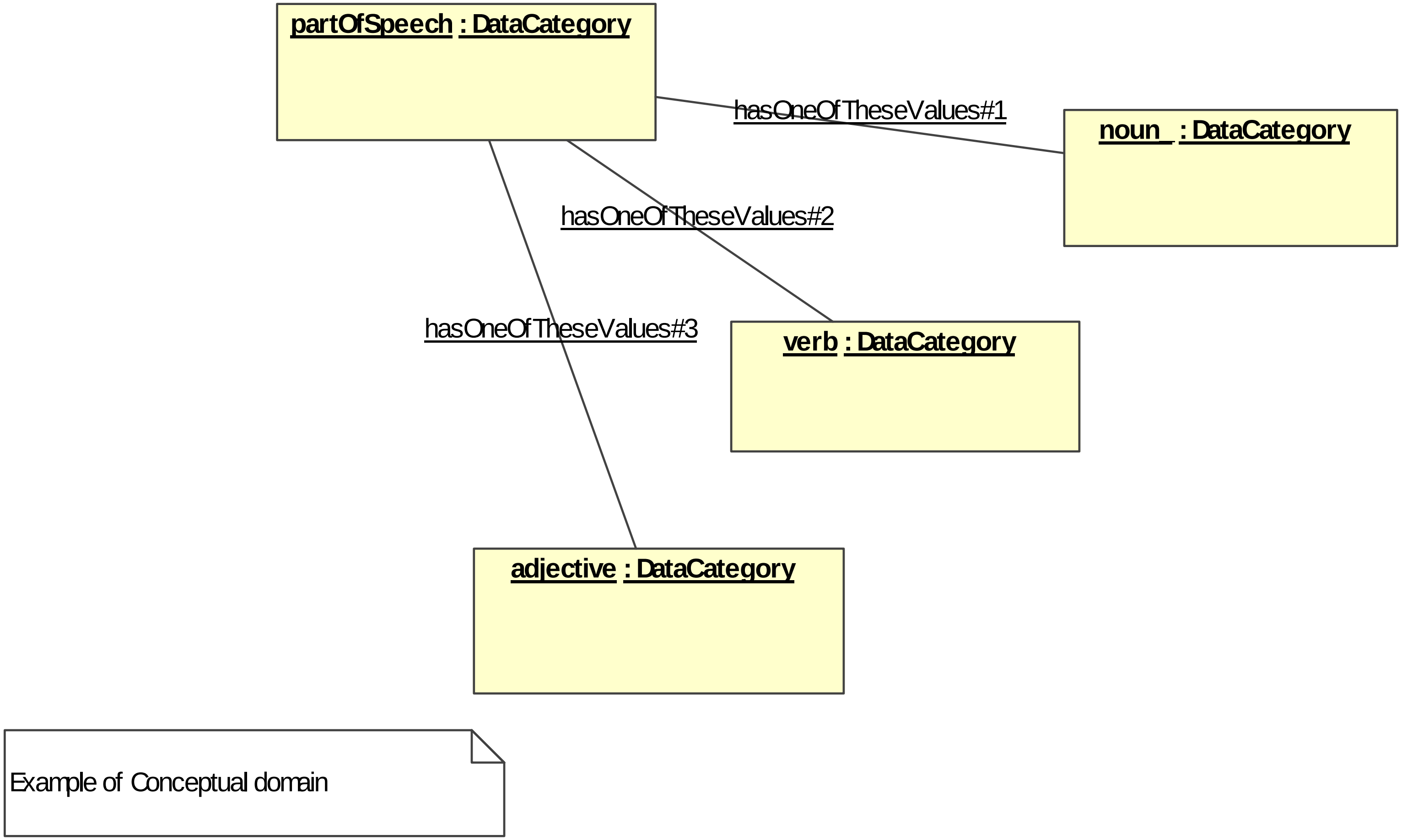


noun : DataCategory

hasABroaderDataCategory

commonNoun : DataCategory

Example of broaderLink


partOfSpeech : DataCategory
hasOneOfTheseValues#1
noun_ : DataCategory
hasOneOfTheseValues#2
hasOneOfTheseValues#3
verb : DataCategory
adjective : DataCategory
Example of Conceptual domain

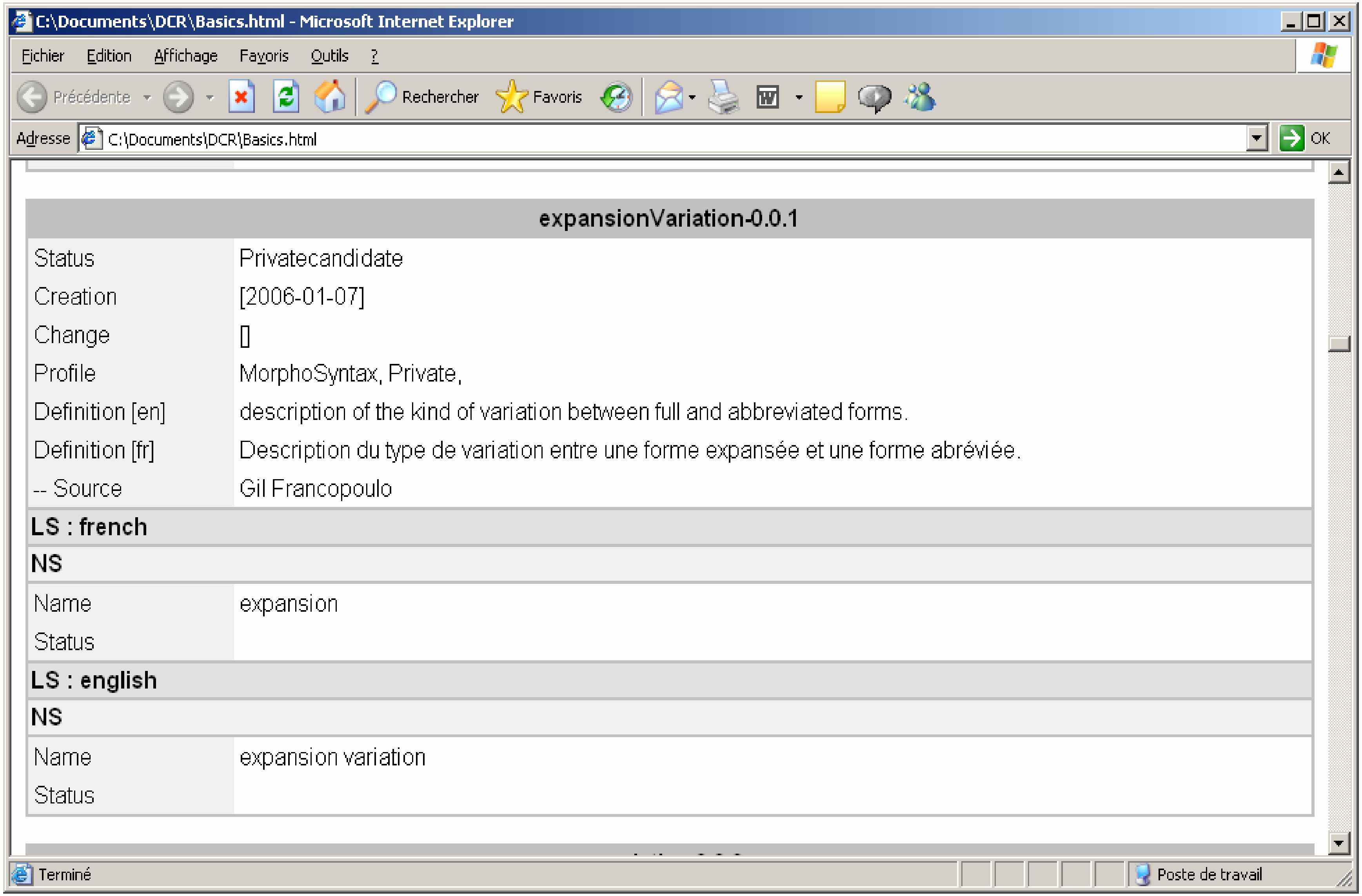
What has been recorded so far in the DCR?

- I use the Syntax software hosted by INIST in Nancy: http://syntax.inist.fr
- The list being rather huge (281 items), the software directories are used in order to help datcat categorization
- PHASE-2: 12 Directories
- In each directory: one or several attribute names and the related values

	#
Basics	36
Cases	35
LanguageTypology	4
MorphemeStemAffix	6
MorphologicalFeaturesExcludingCases	29
Operations	8
PartOfSpeech	78
Reference	6
SemanticallyMotivated	13
SyntacticallyMotivated	40
TransliterationTranscription	7
Dont’tKnowWhatToDo	19

Basics (extract)

- abbreviation
  comment
  derivation
  elision
  expansionVariation
  foreignText
  homograph
  label
  native
  spokenForm
  writtenForm



expansionVariation-0.0.1	
Status	Privatecandidate
Creation	[2006-01-07]
Change	[]
Profile	MorphoSyntax, Private,
Definition [en]	description of the kind of variation between full and abbreviated forms.
Definition [fr]	Description du type de variation entre une forme expansée et une forme abréviée.
-- Source	Gil Francopoulo
LS : french	
NS	
Name	expansion
Status	
LS : english	
NS	
Name	expansion variation
Status	

Cases

- att=case
- val=accusativeCase, dativeCase, genitiveCase etc.

LanguageTypology

- att=languageTypology
- val=agglutinating, inflectional, isolating

MorphemeStemAffix

- affix, infix, morpheme, prefix, stem, suffix

Operations

- addAfter, addBefore, copy etc.

Reference

- Anaphora, antecedent, cataphora, coreference, endophora, referent

MorphologicalFeaturesExclCases

- Attributes like grammaticalGender, mood, tense etc.
- Values like feminine, indicative, present etc.

TransliterationTranscription

- Transliteration, romanization, transciption, script

PartOfSpeech

- Attribute=partOfSpeech
- A small hierarchy of values in order to provide two levels of detail:
  - commonNoun vs noun
  - preposition vs adposition

SyntacticallyMotivated

- Attributes like function, voice
- Values like subject, activeVoice

SemanticallyMotivated

- agent, collective, intensive
- If present in TDG3, will be deleted

Don’tKnowWhatToDo

- invertedComma, colon, instructive etc.

What is left to be done?

- the organization in directories seems fine

=> every definition must be checked: some are rather fuzzy or incorrect

=> We have to check whether the values cover correctly:

- word segmentation and annotation
- Asian, African and Semitic languages

-I just asked the Italian+German team for an informal meeting in Paris dedicated to partOfSpeech and Morphological Features (because the italian team could not come in Jeju). Feel free to come.

And I have some questions:
* Did you consult the morpho-syntax profile?
* Do you think that some particular constants could have been forgotten?
* What is your opinion about this work?

Thank you