PASSAGE Syntactic Representation: a Minimal Common Ground for Evaluation

A. Vilnat (LIMSI & Univ. Paris-Sud), P. Paroubek (LIMSI), E. de la Clergerie (Alpage-INRIA), G. Francopoulo (Tagmatica), M.L. Guénot (Univ. Paris 4)

May 20, 2010
1. General presentation

2. Linguistic phenomena
   - Syntax vs. Semantics
   - Subject relation
   - Coordination

3. Standard XML format

4. Conclusion and Perspective
### What is PASSAGE

PASSAGE (ANR-06-MDCA-013): 
*Produire des annotations syntaxiques à grande échelle*  
(Large Scale Production of Syntactic Annotations)

### Main tasks

- annotating a French corpus of about 100 million words using 10 parsers;
- manually building an annotated reference (400,000 words);
- merging the resulting annotations in order to improve annotation quality;
- performing knowledge acquisition from combined annotations;
- running two parsing evaluation campaigns.
Context: PASSAGE syntactic annotation

6 kinds of syntactic groups (small, generally not embedded,...), 14 syntactic relations linking groups and/or word forms.
### Context: How to compare this annotated corpus?

#### Why this annotation?
- to allow different parsing approaches (from shallow to deep)
- to retrieve a syntactic dependency structure
- with a possible matching from the results obtained by (at least) 10 parsers...

#### Questions
- is it sufficient to deal with most linguistic phenomena?
- does it constitute a sufficient ground to go further (semantics)?
- is it possible to compare/link it with other annotation formalisms?
Syntactic head vs. Semantic head

Some examples

- [le président]$_{GN1}$ [des États-Unis]$_{GP2}$
  - *president of the United States*

- [en guise]$_{GP1}$ [de récompense]$_{GP2}$
  - *by way of reward*

- [cet imbécile]$_{GN1}$ [de Pierre]$_{GP2}$
  - *this fool Pierre*

→ same syntactic head: MOD-N(GP2,GN1)
→ different semantic heads: *président, récompense, Pierre*
Syntax vs. Semantics: Valency vs. Transitivity

Some examples

- \([\text{Je mange}]_{NV1} [\text{de la soupe}]_{GN2}\) *I am eating soup*
  Relations: SUJ-V(Je, mange), COD-V(GN2, NV1)
  Valency (argument structure): *manger* (*je, soupe*)
  → Identical structures

- \([\text{Il mange}]_{NV1} \text{mais } [\text{ne grossit}]_{NV2} [\text{pas}]_{GR3}\) *He eats (a lot) but does not become fat*
  Relations: SUJ-V(II, mange), no COD-V
  Valency (argument structure): *manger* (*il, ∅*)

→ PASSAGE does not annotate the lack of a relation which is semantically expected but syntactically not realised.
Syntax vs. Semantics: Valency vs. Transitivity

Example 1

[Le vent]_{GN1} [souffle]_{NV2}

*The wind is blowing*

Relations: SUJ-V(GN1, NV2)

Valency (argument structure): *souffler (vent)*

→ Identical structures: the subject is the first semantic argument

Example 2

[Il souffle]_{NV1} [un vent]_{GN2} [à décorner]_{PV3}[les bœufs]_{GN4}

*It is blowing a gale*

Relations: SUJ-V(Il, souffle), COD-V(GN2, NV1),...

Valency (argument structure): *souffler (un vent)*

→ the COD-V is the first argument
Subject relation: Control

Infinitive

- [Pierre]$_{GN1}$ [propose]$_{NV2}$ [à Paul]$_{GP3}$ [de venir]$_{PV4}$
  
  *Pierre proposes Paul to come*

  Relations: SUJ-V(GN1, NV2), SUJ-V(GP3, PV4)

- [Avant de partir]$_{PV1}$ [Marie]$_{GN2}$ [éteint]$_{NV3}$ [la lumière]$_{GN4}$
  
  *Before leaving, Marie switches off the light*

  Relations: SUJ-V(GN2, NV3), SUJ-V(GN2, PV1)

- [Fumer]$_{NV1}$ [tue]$_{NV2}$
  
  *Smoke kills*

  Relations: SUJ-V(NV1, NV2)

  → The verb *fumer* has no subject
For a long time, I have lived as they do, and I suffered the same illness.

→ SUJ-V : agreement constraint

→ SUJ-V + AUX-V gives the subject of the main verb.
Subject relation: Passive

Infinitive

- [Pierre]\textsubscript{GN1} [est]\textsubscript{NV2} [applaudi]\textsubscript{NV3}
  *Pierre is applauded*
  Relations: SUJ-V(GN1, NV2), AUX-V(NV2, NV3)
  → The verb *applaudi* has no deep subject.

- [Le livre]\textsubscript{GN1} [est]\textsubscript{NV2} [applaudi]\textsubscript{NV3} [par la critique]\textsubscript{GP4}
  *The book is applauded by critics*
  Relations: SUJ-V(GN1, NV2), AUX-V(NV2, NV3), CPL-V(GP4, NV3)
  → The verb *applaudi* has a deep subject annotated as CPL-V.
Coordination: 3 annotations

SD and GR annotations come from (Marneffe & Manning 08)
Specifications and requirements

- ISO TC37 specifications for morpho-syntactic and syntactic annotation:
  - MAF (ISO 24611)
  - SynAF (ISO 24615)

- The format used during the previous EASY campaign in order to minimize porting effort
- The degree of legibility of the XML tagging.
Standard XML format

Figure: UML diagram of the structure of an annotated document
Les chaises
Conclusion and perspective

Open questions

- is it sufficient to deal with some well known linguistic phenomena?
  → for our main goal (syntactic features): an experimental proof ...

- does it constitute a sufficient ground to go further (semantics)?
  → we hope so! At least, we have the necessary information to do it

- is it possible to compare/link it with other annotation formalisms?
  → Just at the beginning...

- new question: how to address other languages?
  → to be studied for specific syntactic features
Conclusion and perspective

Perspective

- to compare our annotation scheme with what is done in Italy, in EVALITA, with TUT and CoNLL formalisms
- an Italian text and a French one (European texts) annotated following the different annotation schemes, with possible projection from each scheme onto the other.
- and with other languages...