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

Outline

Context : PASSAGE project

What is PASSAGE

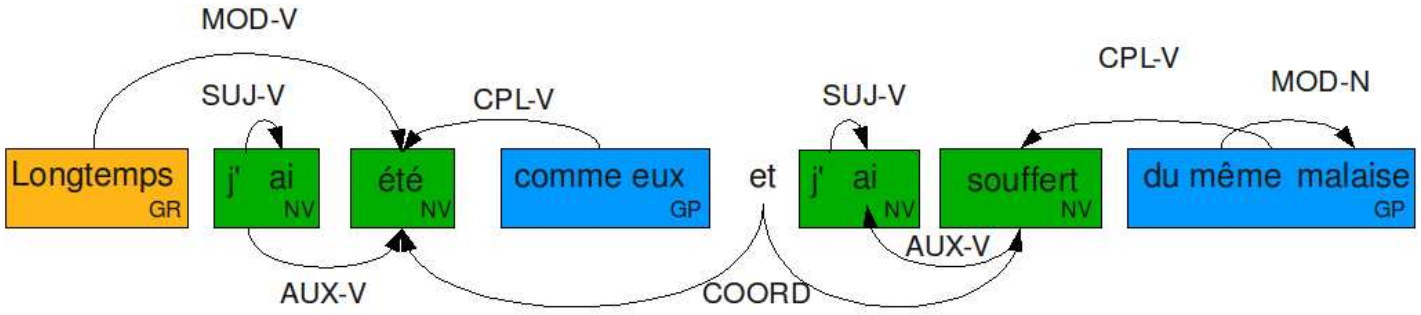
PASSAGE (ANR-06-MDCA-013):
Produire des annotations syntaxiques à grande échelle
(Large Scale Production of Syntactic Annotations)

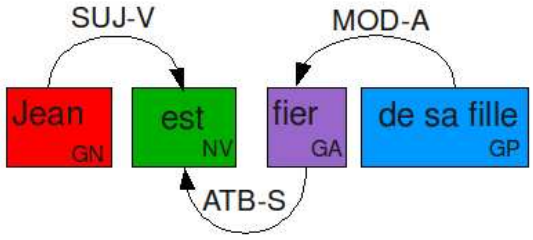
Main tasks

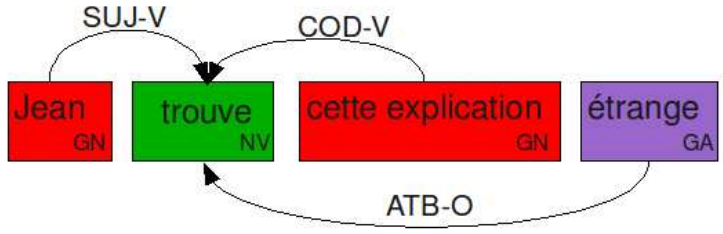
- annotating a French corpus of about 100 million words using 10 parsers;
- manually building an annotated reference (400,000 words);
- merging the resulting annotations in order to improve annotation quality;
- performing knowledge acquisition from combined annotations;
- running two parsing evaluation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

$\longrightarrow$ same syntactic head: MOD-N(GP2,GN1)
$\longrightarrow$ different semantic heads: *président*, *récompense*, *Pierre*

Syntax vs. Semantics: Valency vs. Transitivity

Some examples

- [Je mange]$_{NV1}$ [de la soupe]$_{GN2}$ *I am eating soup*
  Relations : SUJ-V(Je, mange), COD-V(GN2, NV1)
  Valency (argument structure) : *manger (je, soupe)*
  $\longrightarrow$ Identical structures
- [Il mange]$_{NV1}$ mais [ne grossit]$_{NV2}$ [pas]$_{GR3}$
  He eats (a lot) but does not become fat
  Relations : SUJ-V(Il, mange), no COD-V
  Valency (argument structure) : *manger (il, $\emptyset$)*

$\longrightarrow$ 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)*

$\longrightarrow$ 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)*

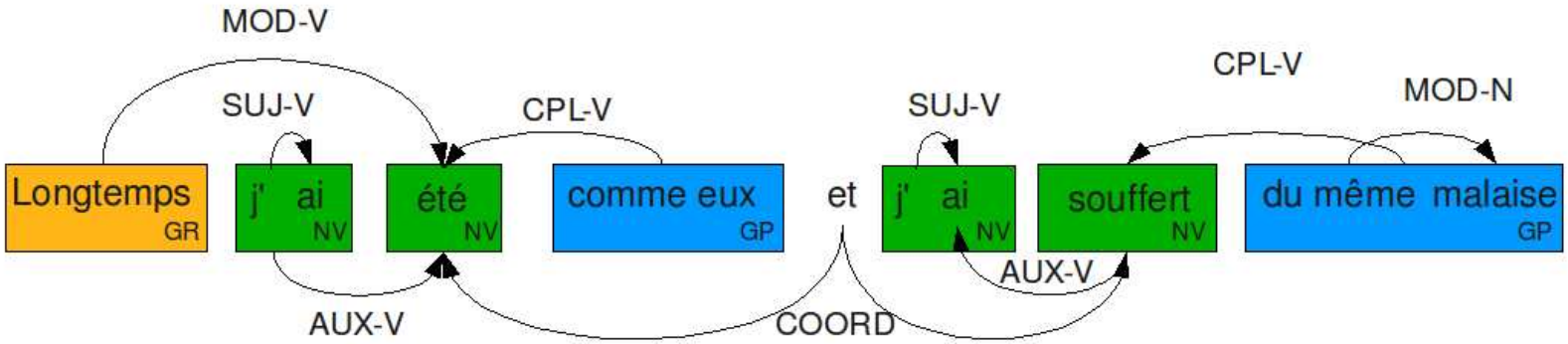
$\longrightarrow$ the COD-V is the first argument

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 swithches off the light
  Relations : SUJ-V(GN2, NV3), SUJ-V(GN2, PV1)
- [Fumer]$_{NV1}$ [tue]$_{NV2}$
  Smoke kills
  Relations : SUJ-V(NV1, NV2)
  $\rightarrow$The verb *fumer* has no subject

Subject relation: compound tenses



For a long time, I have lived as they do, and I suffered the same illness

$\longrightarrow$ SUJ-V : agreement constraint

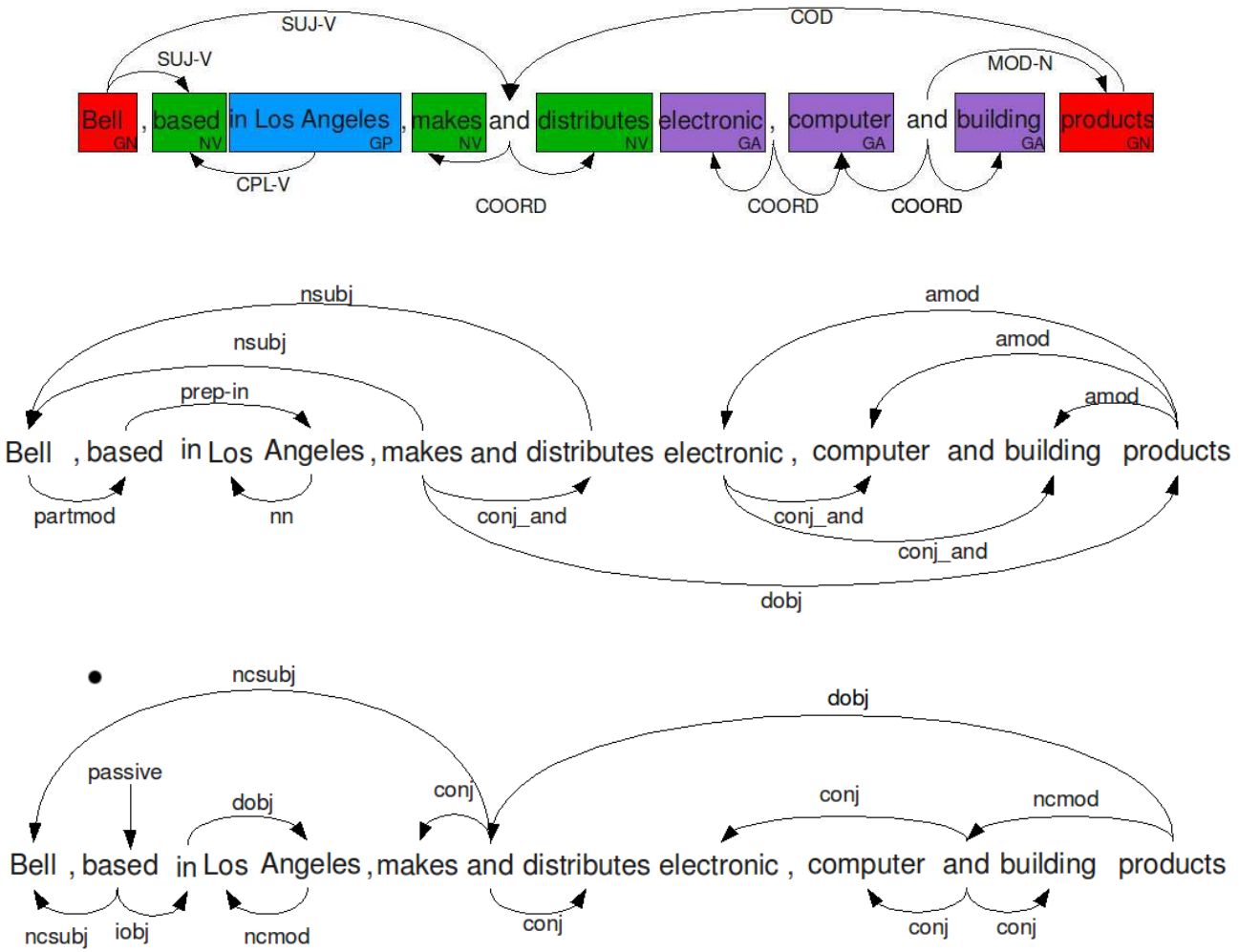
$\longrightarrow$ SUJ-V + AUX-V gives the subject of the main verb.

Subject relation : Passive

Infinitive

- [Pierre]$_{GN1}$ [est]$_{NV2}$ [applaudi]$_{NV3}$
  Pierre is applaused
  Relations : SUJ-V(GN1, NV2), AUX-V(NV2, NV3)
  $\longrightarrow$The verb *applaudi* has no deep subject.
- [Le livre]$_{GN1}$ [est]$_{NV2}$ [applaudi]$_{NV3}$ [par la critique]$_{GP4}$
  The book is applaused by critics
  Relations : SUJ-V(GN1, NV2), AUX-V(NV2, NV3), CPL-V(GP4, NV3)
  $\longrightarrow$The verb *applaudi* has a deep subject annotated as CPL-V.

Coordination: 3 annotations



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

Standard XML format

Specifications and requirements

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

    `http://lirics.loria.fr/doc_pub/maf.pdf`
  - SynAF (ISO 24615)

    `http://lirics.loria.fr/doc_pub/N421_SynAF_CD_ISO_24615.pdf`
- 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

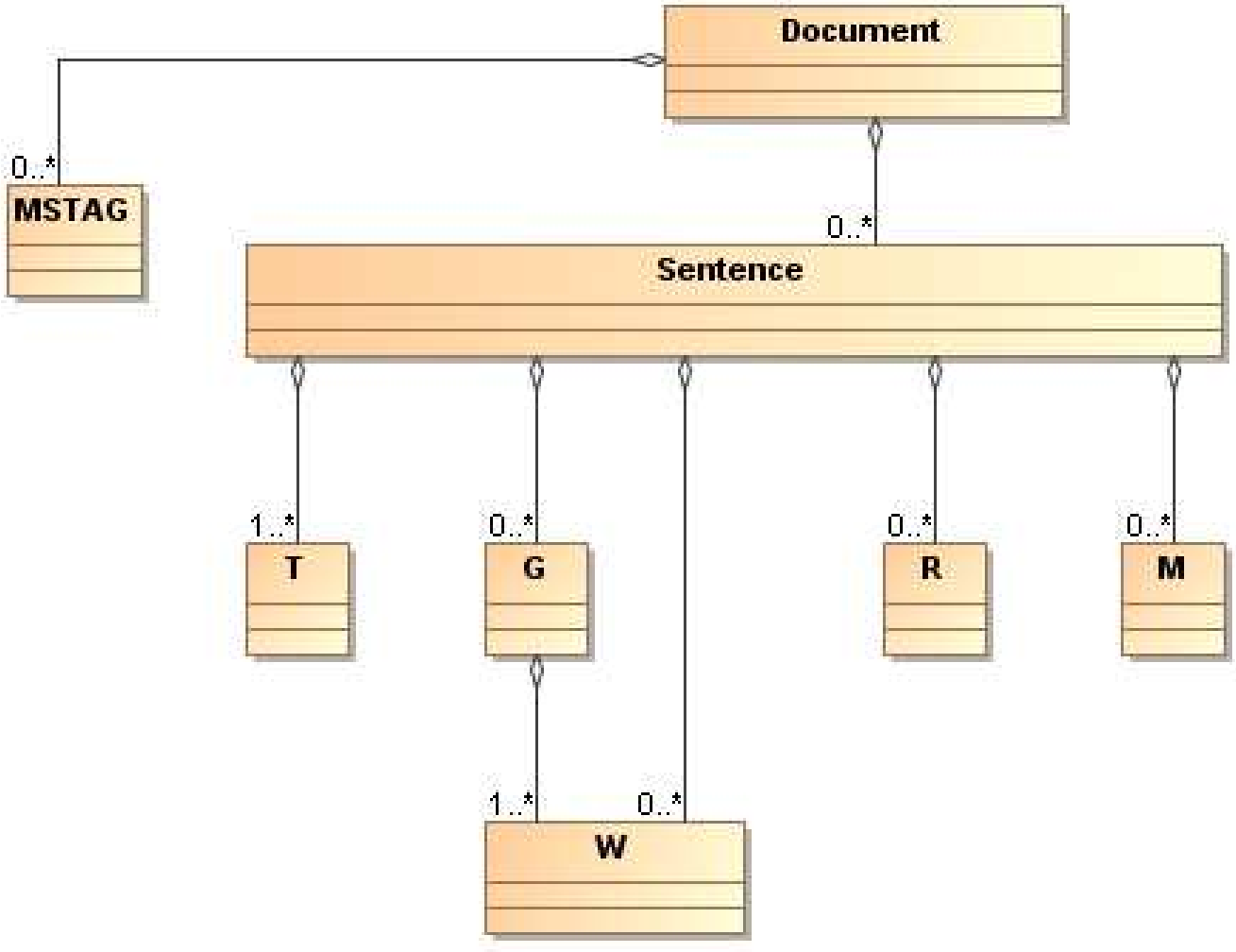


Figure: UML diagram of the structure of an annotated document

Standard XML format

```
<T id="t0" start="0" end="3"> Les </T>
<W id="w0" tokens="t0"
    pos="definiteArticle"
    lemma="le"
    form="les"
    mstag="nP" />
<T id="t1" start="4" end="11"> chaises </T>
<W id="w1" tokens="t1"
    pos="commonNoun"
    lemma="chaise"
    form="chaises"
    mstag="nP gF" />
```

Conclusion and perspective

Open questions

- is it sufficient to deal with some well known linguistic phenomena?
  $\longrightarrow$ for our main goal (syntactic features): an experimental proof ...
- does it constitute a sufficient ground to go further (semantics)?
  $\longrightarrow$ we hope so! At least, we have the necessary information to do it
- is it possible to compare/link it with other annotation formalisms? $\longrightarrow$ Just at the beginning...
- new question: how to address other languages?
  $\longrightarrow$ 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 frm each shema onto the other.
- and with other languages...