Automatic annotations : the syntactic information

Stéphane Rauzy

Laboratoire Parole et Langage
CNRS & Université de Provence

stephane.rauzy@lpl-aix.fr

Objectives

In input, a transcription of spontaneous speech in interaction (e.g. CID copora)

In output, the associated syntactic information :

- Part-Of-Speech tagging, e.g. `Det`, `Noun`, ...
- Groups and syntactic tree structure, e.g. `NP`, `VP`, ...
- Functional relations, e.g. `SUB`, `OBJ`, ...

Strategy :

- Apply the existing tools developed for written textual input
- Adapt them for the treatment of speech transcription
- Investigate and manually correct the annotations in output
- Propose a new model for treating spontaneous speech

LPL tagger, chunker, and parser for written french

The LPL resources and tools for the syntactic treatment of french textual entries (see for example Rauzy & Blache 2009) :

- Lexicon and tokenizer
- Stochastic tagger
- Stochastic chunker and deep parser

Based on Patterns model approach (stochastic model), the grammar is learned/extracted from an annoted corpus. The quality of the tools depends on :

- The size and the coverage of the training corpus
- The quality of the Gold Standard annotations

LPL tagger, chunker, and parser for written french

Stochastic models (e.g. Patterns model) allows to associate a probability to any sequence of categories (e.g. POS tags), and thus describes the regularities found in their distribution.

Training stage :

- Learn the parameters of the model on an annotated corpus, e.g. 853 occurrences of `Noun Verb Det Adj`, followed by :

pattern context	category	occurrences	proba
`Noun Verb Det Adj`	`Pct`	12	0.015
	`Coord`	34	0.045
	`Noun`	807	0.94

Automatic annotation stage :

- Apply the model on raw data, the solution is the one maximizing the overall probability of the sequence.

Part-Of-Speech tagging

A lexicon allows to associate to each form (Part-of-Speech) of the sentence its corresponding lexical tags distribution, e.g.

form	lemma	sampa	tag	frequency
est	être	E	Aux	1671
est	être	E	Verb	21395
est	est	Est	Noun	422

The tagger operates the desambiguation process by choosing the most probable solution, e.g.

Sentence :	La	valise	est	dans	le	coffre	.
Propositions :	Det	Noun	Noun	Noun	Det	Noun	Pct
	Noun		Verb	Prep	Pro	Verb	
	Pro		Aux				

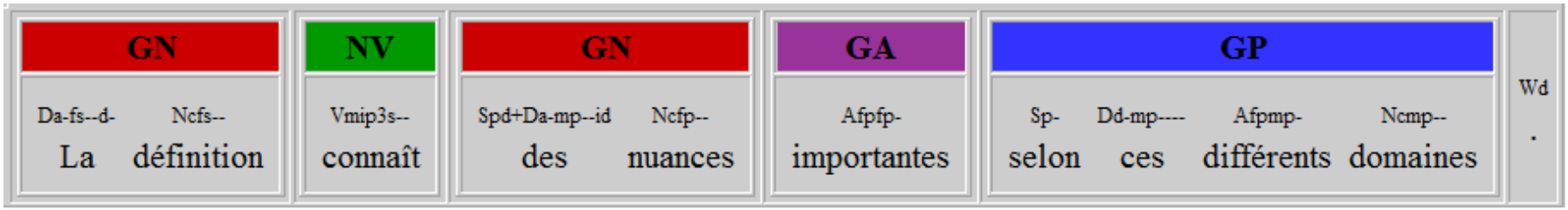
Maximal probability : Det Noun Verb Prep Det Noun Pct

Part-Of-Speech tagging

Form	Solution	Score	A-score	Propositions				
La	Da-fs--d-	-6.6362705	A	Pp3fsj-	Da-fs--d-	Ncm---		
définition	Ncfs--	-6.4358025	A	Ncfs--				
connaît	Vmip3s--	0.74682426	C	Vmip3s--				
des	Spd+Da-mp--id	-4.5408936	A	Da-mp--i-	Sp-+Da-fp--dd	Spd+Da-fp--id	Spd+Da-mp--dd	Spd+Da-mp--id
nuances	Ncfp--	-5.697094	A	Ncfp--	Vmip2s--	Vmsp2s--		
importantes	Afpfp-	-1.0718307	A	Afpfp-				
selon	Sp-	-3.0093784	A	Sp-				
ces	Dd-mp----	-0.7771969	A	Dd-fp----	Dd-mp----			
différents	Afpmp-	-0.046440125	A	Afpmp-	Ai-mp-	Di-mp----		
domaines	Ncmp--	-5.917076	A	Ncmp--				
.	Wd	-1.1673737	A	Wd				

- LPL french lexicon : 595.000 entries with frequency computed from a 140 Megawords corpus (mainly newspapers).
- LPL french tagger : Trained on the 700.000 manually tag corrected LPL-Grace corpus. Tags set of 51 categories, score of 0.975 (F-Measure) on written texts.

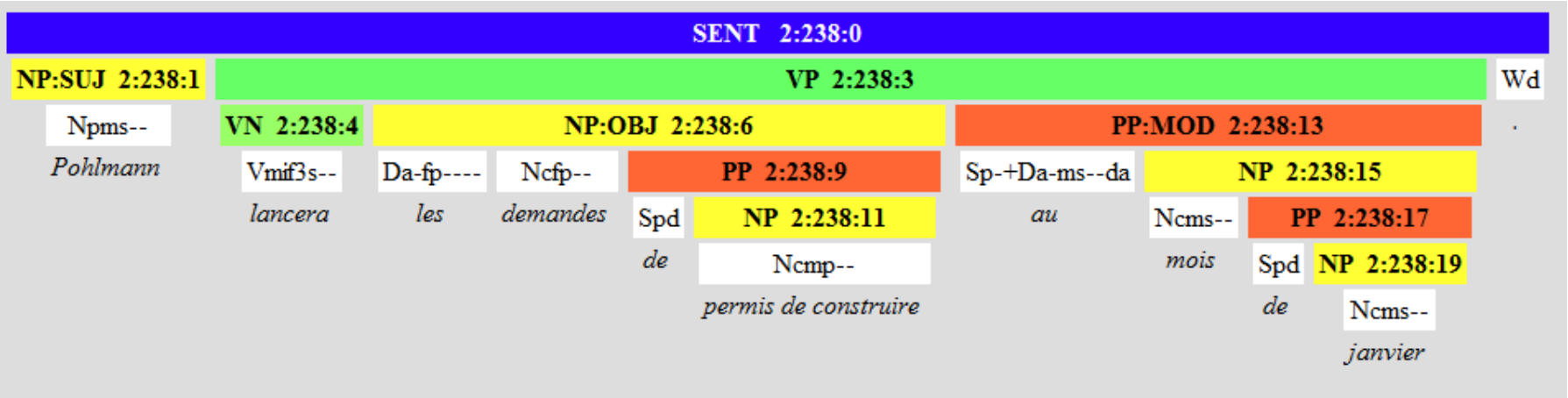
Chunking



Objective : Insert frontiers and label chunks to form syntactic constituents with flat structure.

- LPL french chunker : Trained on the 100.000 Easy gold standard. Easy grammar of seven constituents (GN, GP, NV, ...). Score of 0.93 (F-Measure) on written texts.

Deep parsing



Objective : Form syntactic constituents and tree structure, and indicate functional relations between the constituents.

- LPL french parser : Trained on the 100.000 words LPL-FT corpus, a grammar of 15 constituents (`NP`, `VP`, `VN`, ...) and 9 relations (`SUB`, `OBJ`, `COORD`, ...). Evaluation in progress.

Speech transcription treatment - Filtering

Remove from the transcription the phenomena which are not found in written french :

- hesitation, pause and filled pause (e.g. heu, #, +, ...)
- laughs (e.g. @)
- truncations (e.g. remp-)

TOE : # *alors moi j'y étais allée déjà je comprends rien à ce qu'ils font* + *mais euh j'y étais allée pour remp- pour remplir la salle* #

Filtered TOE : *alors moi j'y étais allée déjà je comprends rien à ce qu'ils font mais j'y étais allée pour pour remplir la salle*

Speech transcription treatment - Punctuation marks

Punctuation marks are not found in the transcription. Allow the tagger to insert these marks based on written french model :

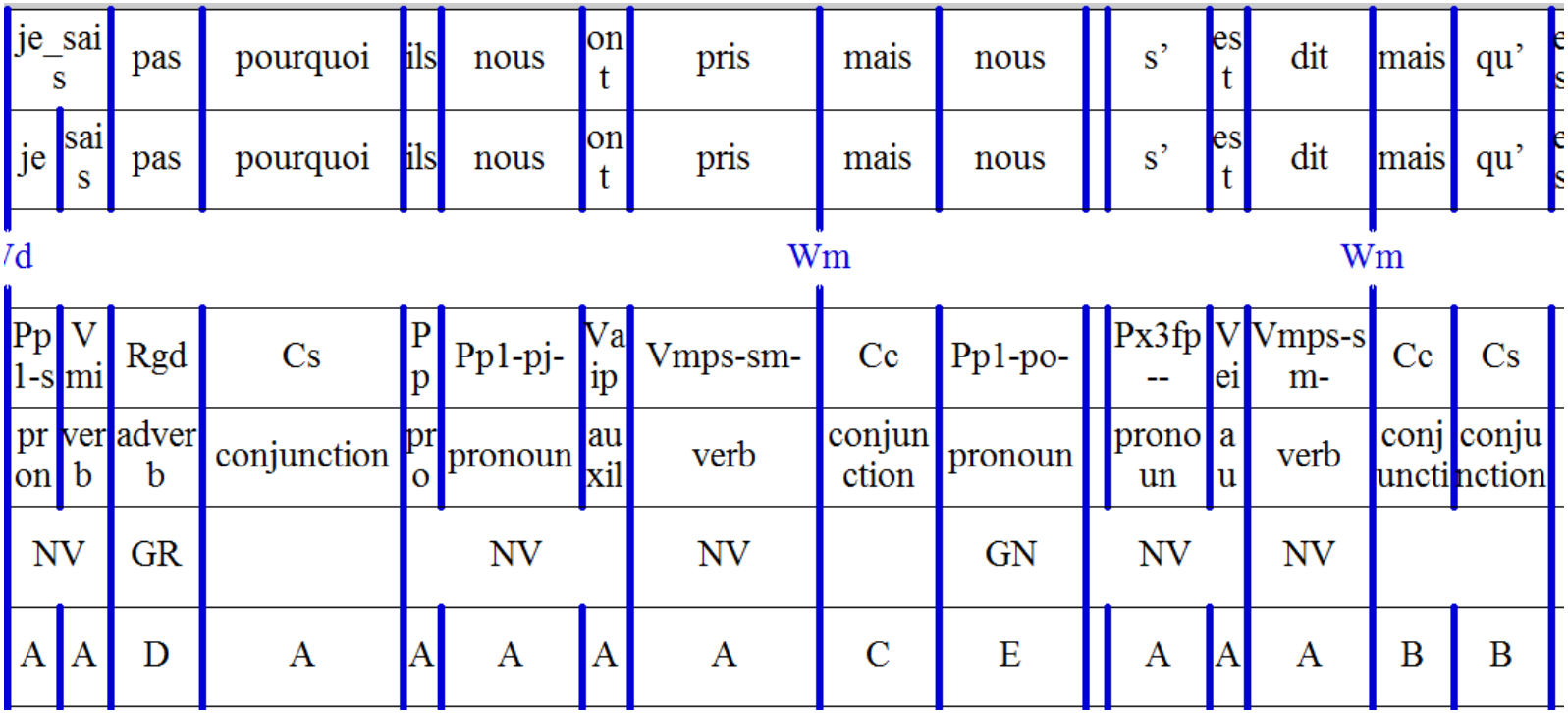
- Wd : Strong punctuations (e.g. ., !) delimiters of sentences
- Wm : Weak punctuations (e.g. ,) delimiters of smaller syntactic units

Example of pattern with punctuation marks insertion :

pattern context	insertion	category	proba
Verb Det Noun	-	SubPro	0.01
	Wd	SubPro	0.34
	Wm	SubPro	0.75

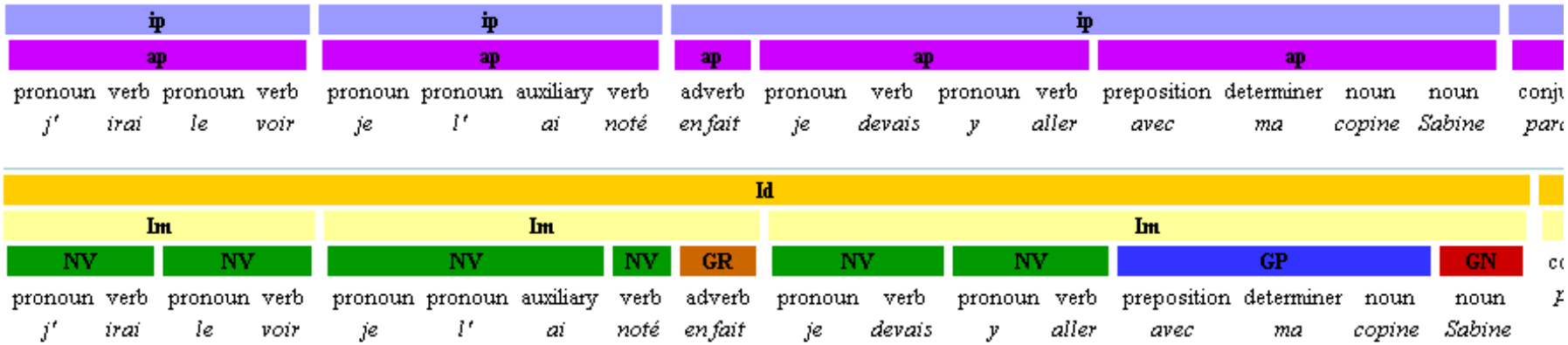
Punctuation marks are inserted if they increase the value of the overall sequence probability.

Pseudo-sentence units



The insertion process allows to identify pseudo-sentence units (and smaller units corresponding to weak punctuations), based on the syntactic content of the transcription.

Pseudo-sentence and prosodic annotation



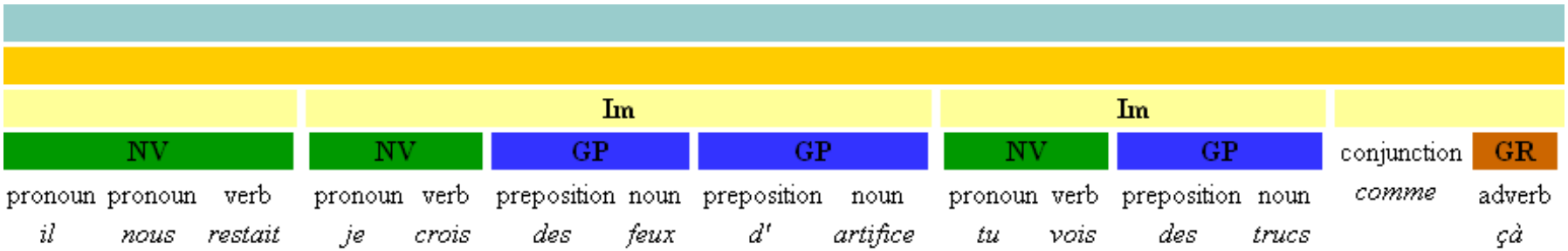
Comparison of the automatic syntactic annotation and manual annotation of prosodic intonational phrase (ip) and accentual phrase (ap). Pseudo-sentence frontiers match ip frontiers at a level of 60% (Nesterenko et al. 2010).

Speech transcription treatment - Lexicon adaptation

Some forms do not play their usual syntactic role. These forms are few but frequent (10 % of the corpus). Their lexical tags distribution has to be modified :

- Add or transform lexical distribution frequencies (e.g. `quoi`, `bon`, `bien`, `putain`, `parce que`, ...

Some other use are more problematic, e.g. `tu vois`, `je crois`, ...



Results and ongoing works

Syntactic annotations for the transcriptions of the CID corpus :

- Tagging and chunking steps are efficient.
- Identification of larger units (pseudo-sentence) corresponding to punctuation marks found in written texts.
- Evaluation of the performance and manual correction in progress.

The deep parser does not work properly (about 15 % of tree formation) :

- Disfluencies do not allow tree structure formation ?
- Not appropriate for the description of speech in interaction ?