MarsaTag, a tagger for French written texts and speech transcriptions

Stéphane Rauzy, Grégoire de Montcheuil & Philippe Blache

Laboratoire Parole et Langage
CNRS & Aix Marseille Université, France

stephane.rauzy@lpl-aix.fr

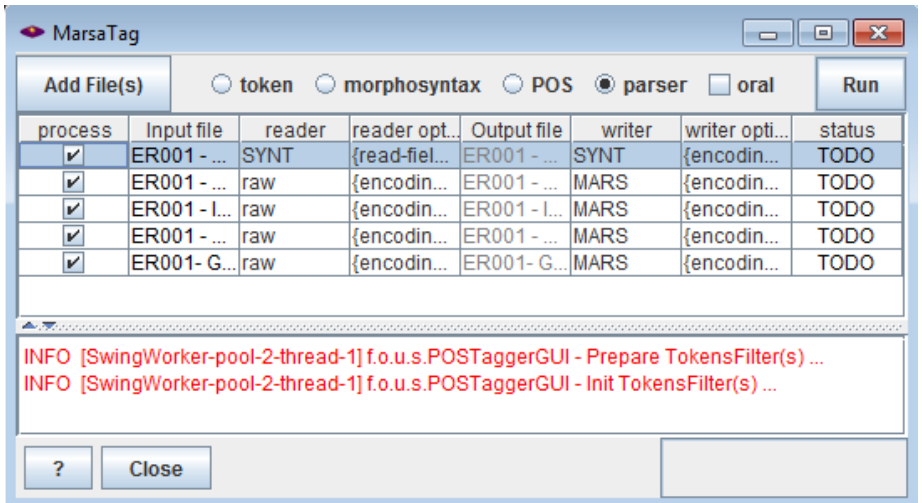
MarsaTag

In input, French productions :

- Written texts
- Speech transcriptions

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`, ...



MarsaTag

(hdl:11041/sldr000841)

The processing chain

The processing chain :

- Pre-lexical treatment, the tokenizer
- Lexical access
- Stochastic tagger
- Stochastic chunker and deep parser

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

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

The tokenizer

Split the sentence in a list of tokens and identify tokens (not in the main lexicon) requiring a special treatment :

- Numbers, dates, ...
- Punctuation marks
- Acronyms, abbreviations, ...
- Proper names

Le budget de l'Armée sera augmenté le 23 août, a indiqué aujourd'hui M. Vasseur à 14H45.

Le|$_L$budget|$_L$de|$_L$l'|$_L$Armée|$_L$sera|$_L$augmenté|$_L$le|$_L$23|$_N$août|$_L$,|$_P$
a|$_L$indiqué|$_L$aujourd'hui|$_L$M.|$_A$Vasseur|$_N$à|$_L$14H45|$_H$.|$_P$

Lexical access

A main lexicon allows to associate to each form of the sentence its corresponding lexical distribution of tags, e.g.

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

At this stage, an ambiguous list of tags is proposed for the tokens of the sentence, e.g.

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

Stochastic grammars, probabilistic models

Probabilistic models (e.g. Patterns model) allows to associate a probability to any sequence of categories (e.g. tags). The model describes the regularities of the stochastic grammar.

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	probability
`Noun Verb Det Adj`	`Pct`	12	0.015
	`Coord`	34	0.045
	`Noun`	807	0.94

Tagging and parsing process :

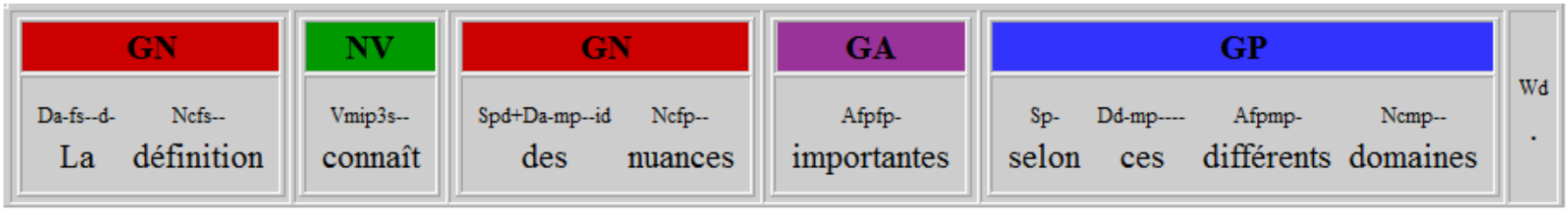
- Apply the model to the ambiguous data input, the best solution is the one maximizing the probability over the sequence.

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				

- MarsaLex lexicon : 595.000 entries with frequency computed from a 140 Megawords corpus (mainly newspapers).
- MarsaTag 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.

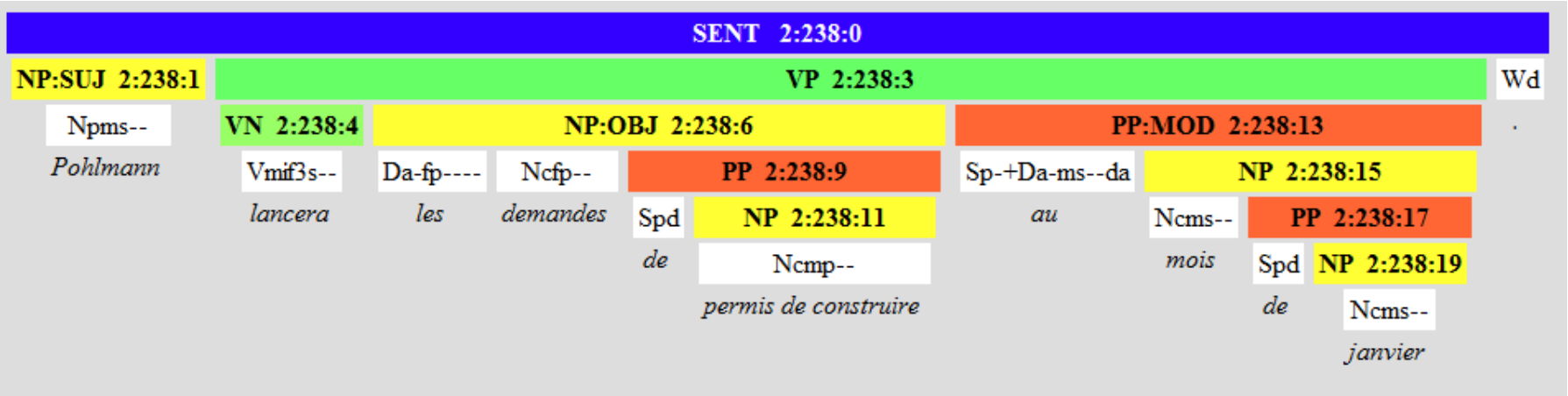
Chunking



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

- MarsaTag 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.

- MarsaTag 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

The tools have been adapted for the treatment of speech transcriptions by investigating the CID corpus (Bertrand et al. 2008)

- CID - Corpus of Interactive Data - 8 hours of spontaneous dialogues in French
- Enriched Orthographic Transcription aligned on signal
- $115,000$ tokens with manually corrected POS-tags
- Disfluencies annotation
- Various others multi-level annotations available...

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, #, +, ...)
- laughter (e.g. @)
- word 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 available in the transcription. The tagger allows to insert these marks based on written french model :

- Wd : Strong punctuations (e.g. ., !) delimiters of sentences
- Wm : Soft 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 probability of the sequence treated.

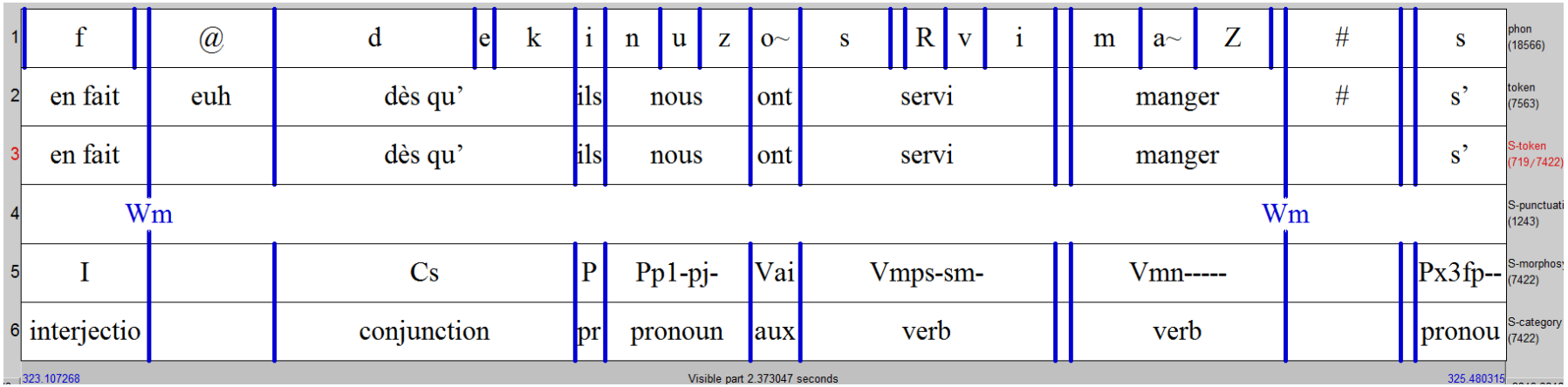
Speech transcription treatment - Lexicon adaptation

- Some forms are used with different and supplementary purposes than in written productions (e.g. spoken discourse markers). These forms are few (less than 40) but frequent (10% of the CID corpus). Their lexical tags distribution have been modified :

form	tag	CID frequency	probability
bon	`Interjection`	634	0.95
bon	`Adverb`	29	0.04
bon	`Adjective`	5	0.01
...	...	...	...

- Forms proper to speech production have been added to the lexicon (word reductions, onomatopoeia, ...)

Speech transcription - Evaluation



MarsaTag speech tagger : an F-measure of 0.947 evaluated on the manually corrected CID corpus.

Chunking looks robust but deep parsing does not work properly :

- Disfluencies disturb the formation of tree structures ?
- Phrase Structure Grammar is not appropriate for modeling the syntax of spontaneous speech in interaction ?

Summary and perspectives

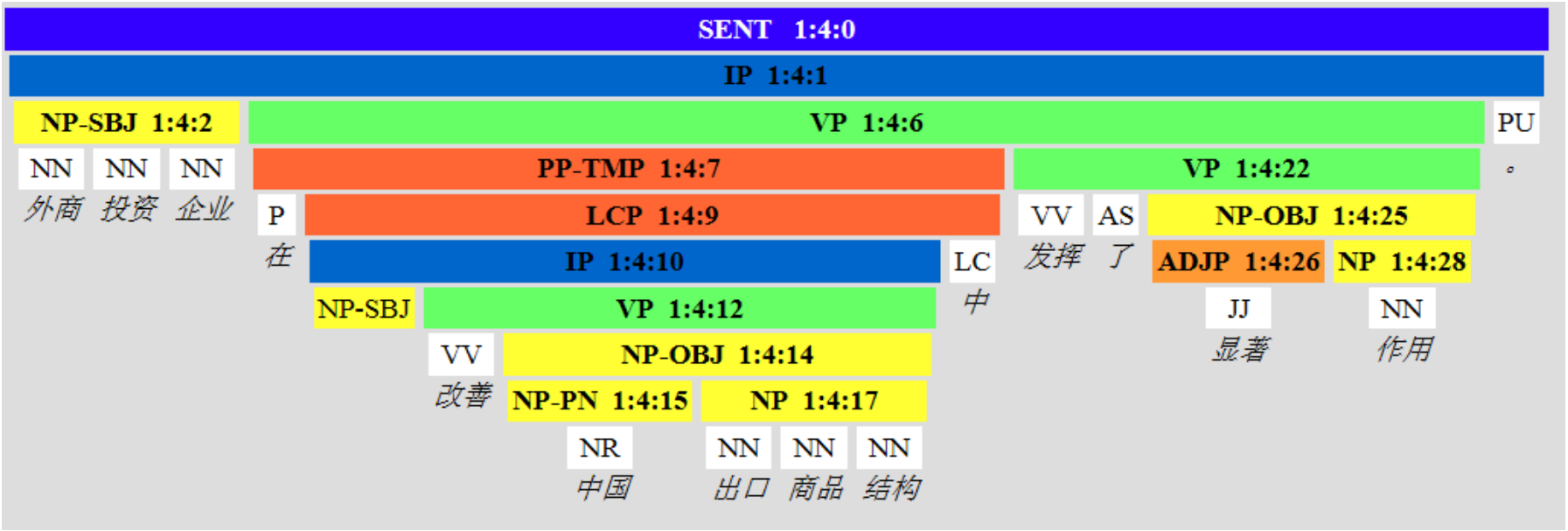
MarsaTag is efficient for extracting the syntactic information from French productions :

- Written texts
- Speech transcriptions (except for deep parsing)

Perspectives :

- Add to the processing chain a model taking into account difluencies phenomena in speech productions.
- Apply our corpus-based approach to other languages with existing resources :
  - e.g. The Chinese Penn Treebank (Xue et al. 2005) for written texts, PolyU Corpus of Spoken Chinese (Yap et al. 2012) for speech transcriptions.

Chinese Penn Treebank



Chinese Mandarin Penn Treebank (Xue et al. 2005) :

- Number of sentences : $51,443$
- Number of tokens : $1,127,832$

Chinese Penn Treebank

Chinese Mandarin Penn Treebank : lexicon, starting with character 也

Form	Frequency	Category	Examples
也	5244	AD	1003:4:8 1003:5:8 1005:8:30 1006:5:25 1007:4:28 1008:2:11 1010:7:21 1013:5:33 1014:6:17 1016:2:63 1016:2:1(
也	77	CC	185:7:31 220:9:32 2249:6:52 239:32:86 242:5:59 245:8:35 250:20:34 258:9:54 282:17:57 291:3:55 359:11:78 460
也	1	NN	2124:0:14
也	30	SP	2232:325:31 2249:6:102 2271:20:8 2271:26:8 2271:28:8 2271:32:8 2271:34:8 2271:50:8 2271:58:8 2271:64:8 227
也可以说	1	AD	2374:112:23
也好	34	AD	2150:238:20 2150:238:30 2150:238:40 2155:333:19 2157:232:52 2157:232:61 2157:253:35 2157:253:62 2225:27(
也好	3	SP	1133:8:26 1133:8:43 2262:15:7
也就是	8	AD	2155:222:8 2156:167:29 2157:177:11 2171:1:40 2292:11:49 2307:18:6 833:20:46 885:46:4
也就是说	29	AD	1557:6:41 1592:2:89 1932:25:35 2156:68:31 2156:245:3 2156:280:71 2156:342:4 2156:360:67 2157:35:4 2157:14
也是	8	AD	2223:9:20 357:5:37 620:3:27 762:13:145 770:2:90 770:80:116 854:69:23 864:25:47
也是	3	CC	739:15:43 797:56:43 809:99:39
也有说	1	AD	2208:2:3
也罢	1	SP	2262:15:14
也罢	1	VV	1352:1:73
也许	107	AD	1108:4:42 1122:2:36 1302:6:12 1336:6:41 1364:12:5 1381:3:63 1457:7:8 1785:10:7 1803:9:6 1829:0:25 1921:10:5
也通社	1	NR	617:6:18
也门	54	NR	1091:0:35 1241:1:20 1241:2:26 1528:3:20 1528:4:7 1533:0:44 1533:1:18 1578:1:30 1578:2:52 1753:33:24 1753:3
也门人	4	NN	1528:3:52 1533:1:52 982:3:52 982:4:44

Lexicon extracted from the Chinese Penn Treebank :

- Number of entries : 71,629