4-Couv : building a new treebank based on backcovers

Grégoire de Montcheuil Philippe Blache Stéphane Rauzy
Marie-Laure Guénot

3rd VariAMU Workshop - Aix-en-Provence - 1st October 2015

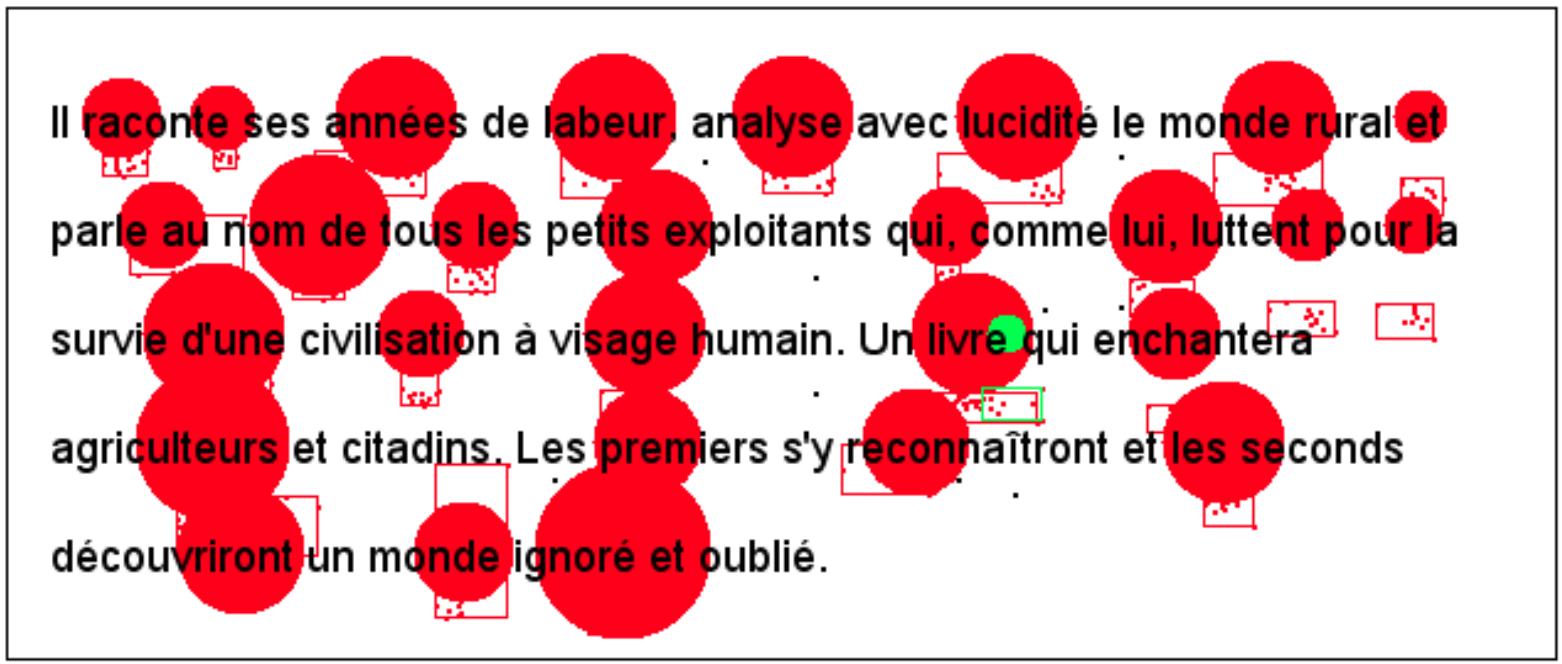
Outline

- Motivation
- The annotation framework
- Treebanking tools

Motivation

A resource for **studying human language processing** :

- e.g. eye-tracking when reading texts



(Demberg and Keller (2008), Rauzy and Blache (2012))

Requirements

Maintain the attention during the reading

- short texts
- semantically consistent
- atemporal
- arousing interest

$\Rightarrow$ text from *backcovers*

Context

Treebanks also essential resource in **linguistic description** and **natural language processing**

$\Rightarrow$ compatibility with the French Treebank (*FTB*, $\sim$ 20.000 trees; Abeillé, Clément, and Toussenel (2003)) and its derivatives (e.g. *MFT*, 3.800 trees; Schluter and Genabith (2007))

- constituency-based
- same lexical categories - richer morphosyntactic features
- same constituent & function tagsets

The annotation framework

- Tokenization
- Syntactic annotation
- Pre-annotation

Tokenization

- **maximal**, i.e. even **highly constrained forms are split**:

  “il était une fois” *(once upon a time)*

  “mettre à nu” *(lay bare)*

- **except** if they **don’t follow syntactic composition rules**:

  “tant mieux” *(even better)*

  “d’autant plus” *(all the more)*

Lexical categories

- 10 categories: Adjective, Determiner, Noun*(common, proper)*, Pronoun*(⊂ clitic)*, Adverb, Preposition, Verb, Connector*(coordinating/subordinating conjunction)*, Interjection, (Punctuation)
  (disused FTB: ~~Prefix~~, ~~Foreign word~~)
- 1-11 features: *(eventually under-specified)*
  - *generic*: **nature**, type, person, gender, number
  - *specific*: e.g.
    (pronoun) grammatical case, reflexive, postposed
    (verb) mood, auxiliary, complements, . . .
- no category shift:

  "une tarte maison" *(an home[made] pie)*

  "il est très zen" *(he is very zen)*

  Noun ~~Adjective~~

Syntactic annotation

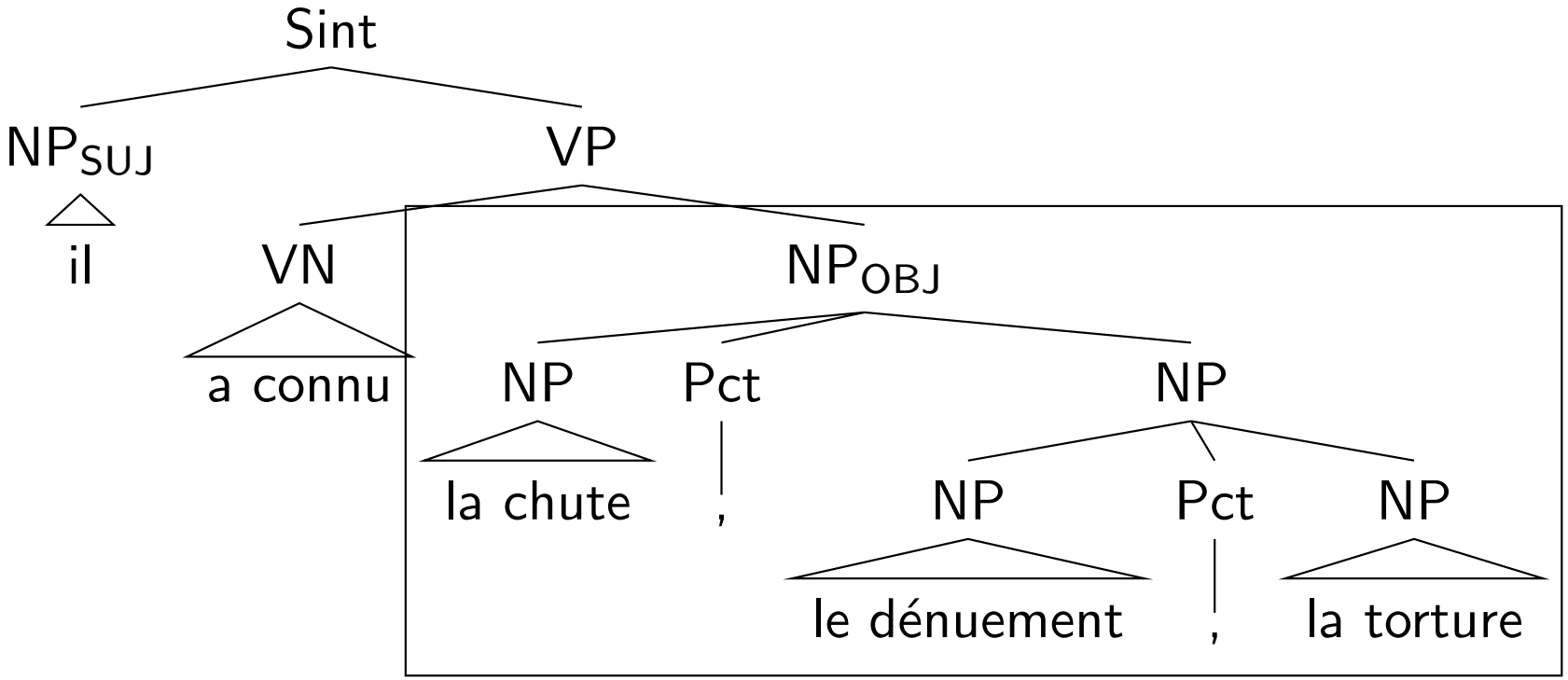
- Only tags corresponding to strictly syntagmatic constructions:
  NP/VP/AP/AdP/PP (noun/verbal/adj./adv./prep. phrase),
  VN (verbal nucleus), VNinf/VNpart (infinitive/participial VN),
  VPinf/VPpart (infinitive/participial clause),
  SENT (sentence), Srel/Ssub/Sint (relative/subordinate/other clause),
  ~~COORD~~ (coordination) *(≠ FTB)*.

- Same syntactic functions:
  SUJ (subject), OBJ (direct object),
  A-OBJ/DE-OBJ/P-OBJ (indirect complement introduced by "à"/"de"/another preposition),
  ATS/ATO (predicative complement of a subject/direct object),
  MOD (modifier or adjunct).

Coordination *(≠ FTB)*

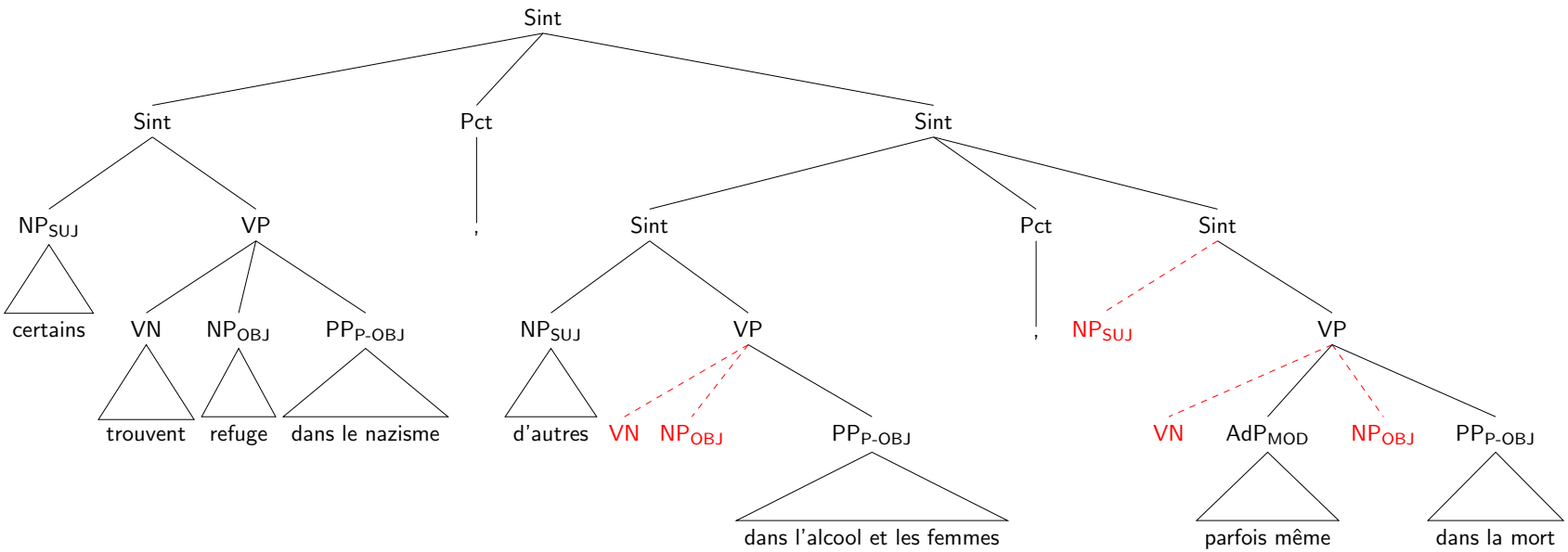
"il a connu la chute, le dénuement, la torture"

(he known the fall, the deprivation, the torture)



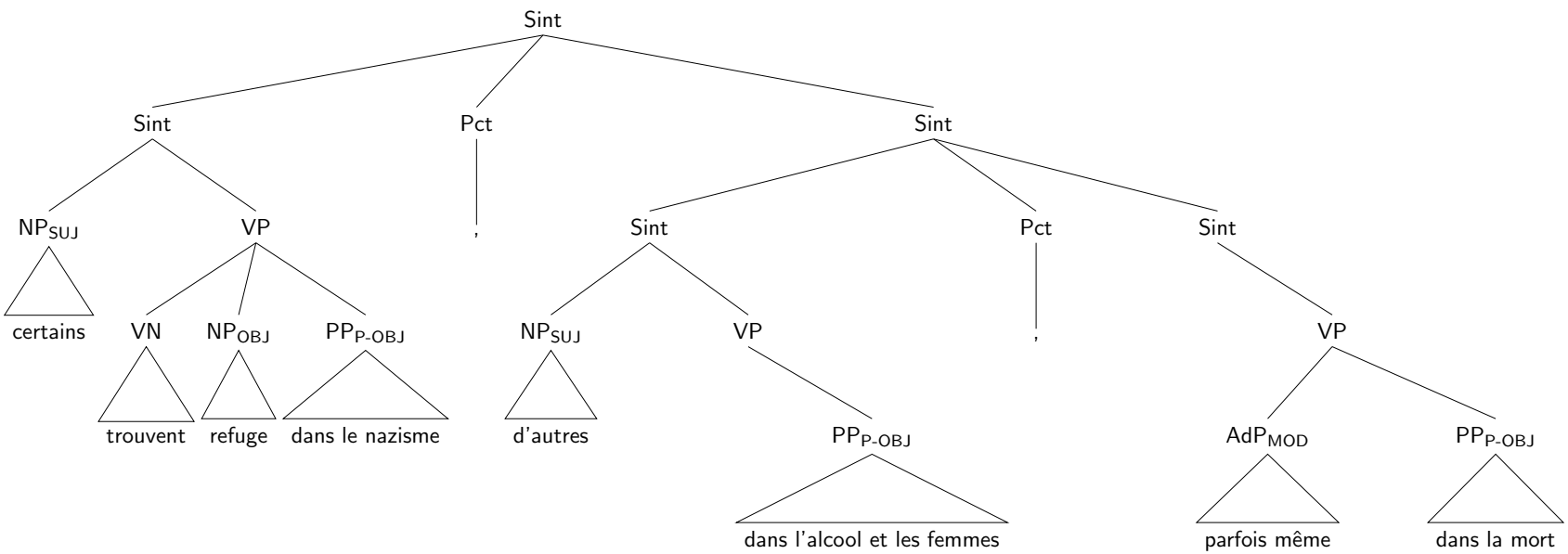
No empty category

- Any empty category is inserted (e.g. elliptical construction)
  "certains trouvent refuge dans le nazisme,
  (some find refuge in the nazism,
  d'autre dans l'alcool et les femmes,
  others in alcohol and women,
  parfois même dans la mort"
  sometimes even in the death)



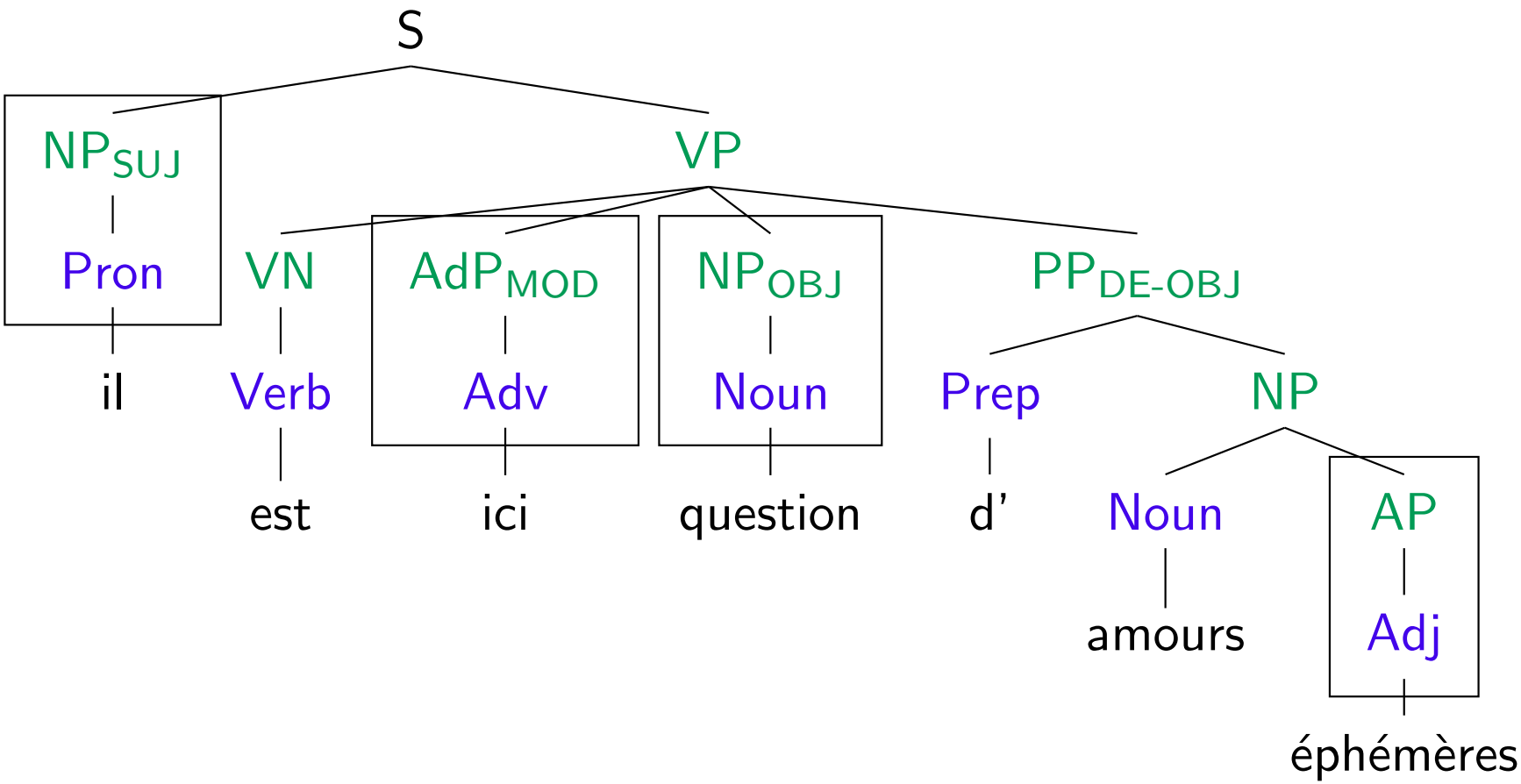
No empty category

- Any empty category is inserted (e.g. elliptical construction)
  "certains trouvent refuge dans le nazisme,
  (some find refuge in the nazism,
  d'autre dans l'alcool et les femmes,
  others in alcohol and women,
  parfois même dans la mort"
  sometimes even in the death)



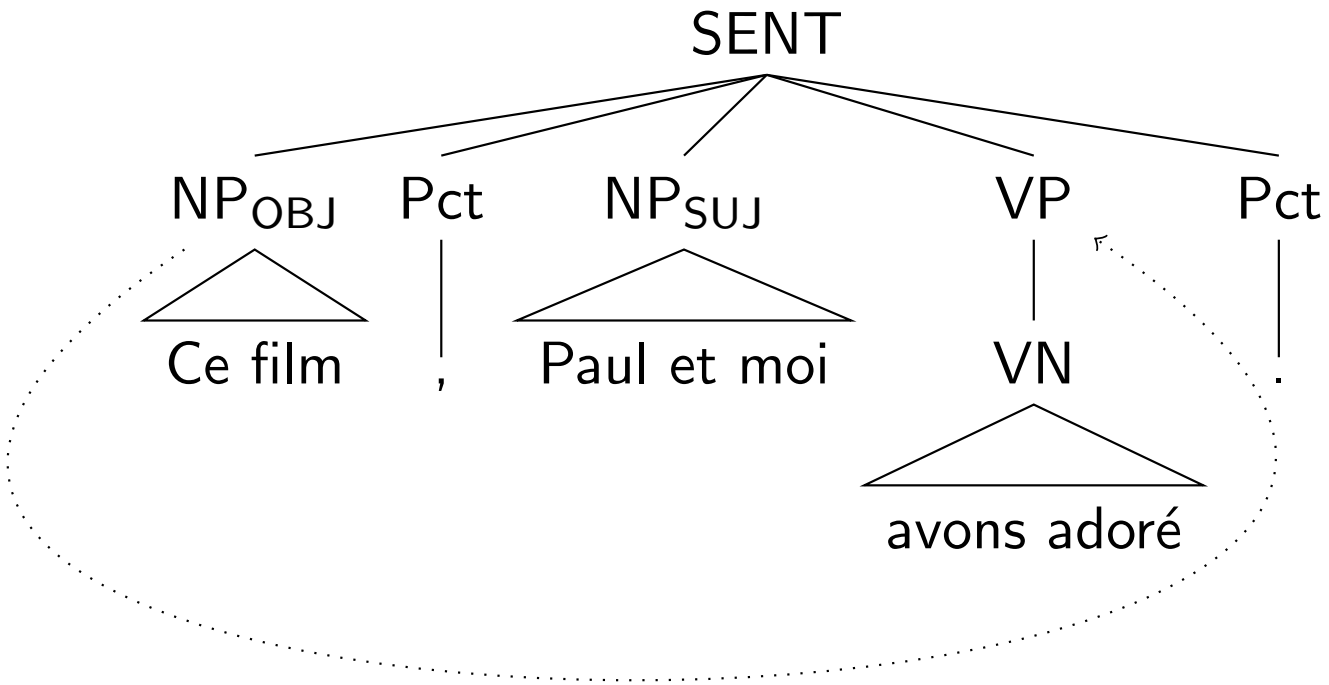
Lexical & syntactic levels

- Distinct lexical & syntactic level ⇒ unary syntagms
  "il est ici question d'amours éphémères"
  (it is here an issue of ephemeral loves)



No discontinuous constituent

- No discontinuous constituent
  "Ce film, Paul et moi avons adoré."
  (This movie, Paul and I really liked.)



Pre-annotation

MarsaLex (hdl:11041/sldr000850)

- French language lexicon
- $\sim$ 595.000 forms, 59.000 lemmas

MarsaTag (hdl:11041/sldr000841)

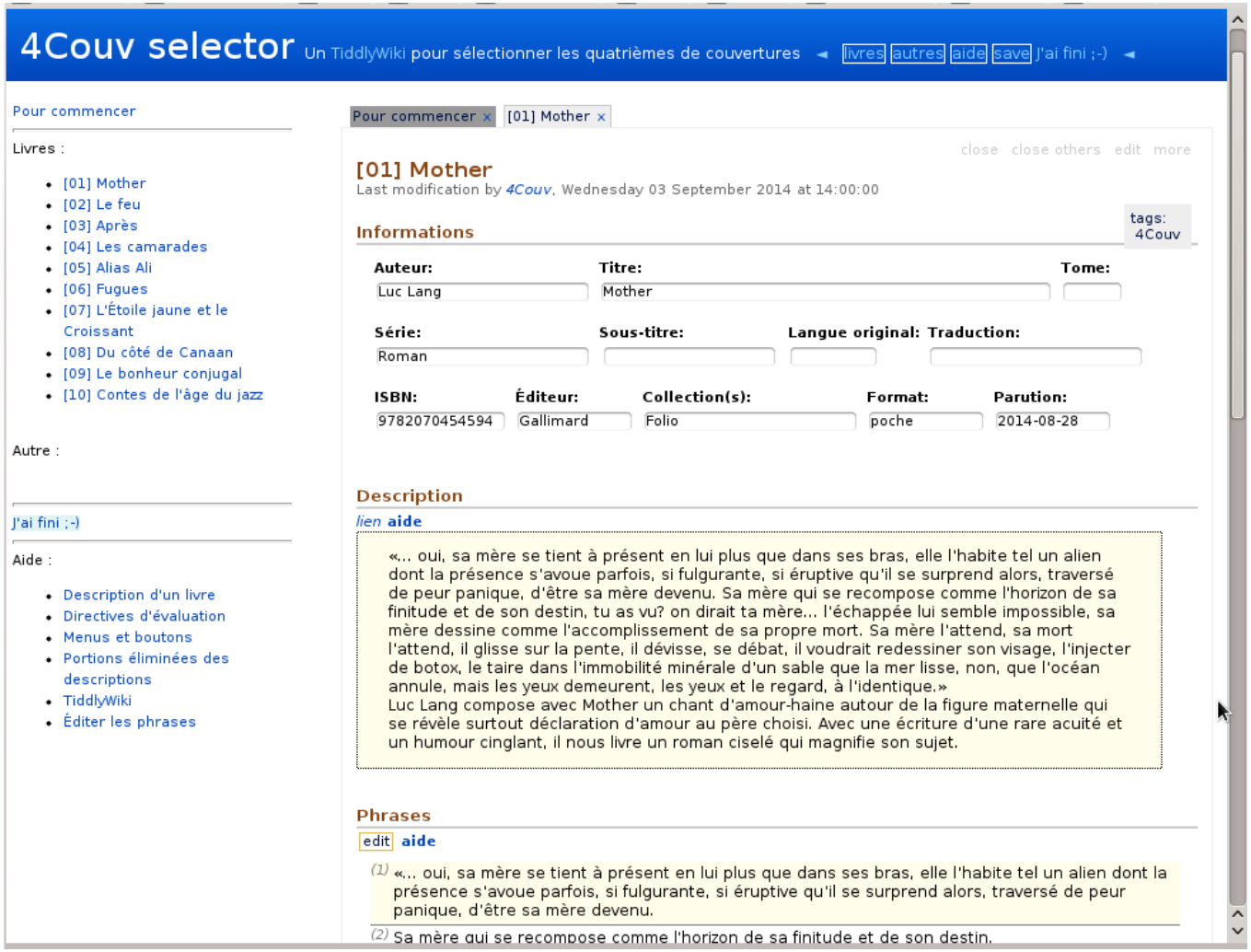
- Stochastic HMM POS tagger & parser (Rabiner (1989))
- POS tagger train on a LPL version of *Grace/Multi-tag* corpus (Paroubek and Rajman (2000)): $\sim$ 700,000 tokens
- Parser train on a LPL version of the *MFT*: 1,500 sentences

Treebanking tools

- Text selection
- Automatic annotation revision
  - morphosyntactic tags
  - constituent trees editor

4Couv selector

- Wiki embedded in autonomous HTML files (TiddlyWiki)
- Presentation of 10 texts to evaluate



Rapid evaluation

Form with boxes and list of choices

Sentence split and sections

Phrases

edit **aide**

«… oui, sa mère se tient à présent en lui plus que dans ses bras, elle l'habite tel un alien dont la présence s'avoue parfois, si fulgurante, si éruptive qu'il se surprend alors, traversé de peur panique, d'être sa mère devenu.
Sa mère qui se recompose comme l'horizon de sa finitude et de son destin,
tu as vu?
on dirait ta mère…
l'échappée lui semble impossible, sa mère dessine comme l'accomplissement de sa propre mort.
Sa mère l'attend, sa mort l'attend, il glisse sur la pente, il dévisse, se débat, il voudrait redessiner son visage, l'injecter de botox, le taire dans l'immobilité minérale d'un sable que la mer lisse, non, que l'océan annule, mais les yeux demeurent, les yeux et le regard, à l'identique.»
Luc Lang compose avec Mother un chant d'amour-haine autour de la figure maternelle qui se révèle surtout déclaration d'amour au père choisi.
Avec une écriture d'une rare acuité et un humour cinglant, il nous livre un roman ciselé qui magnifie son sujet.

Sentence split and sections

Wiki syntax :

- sentences are table rows
- sections separated by blank lines

Review unknown words

Pour commencer × | [01] Le cycle d'Elric × | Mots inconnus ×

close close others edit more

Mots inconnus

Last modification by *4Couv*, Wednesday 03 September 2014 at 14:00:00

tags: MenuMore

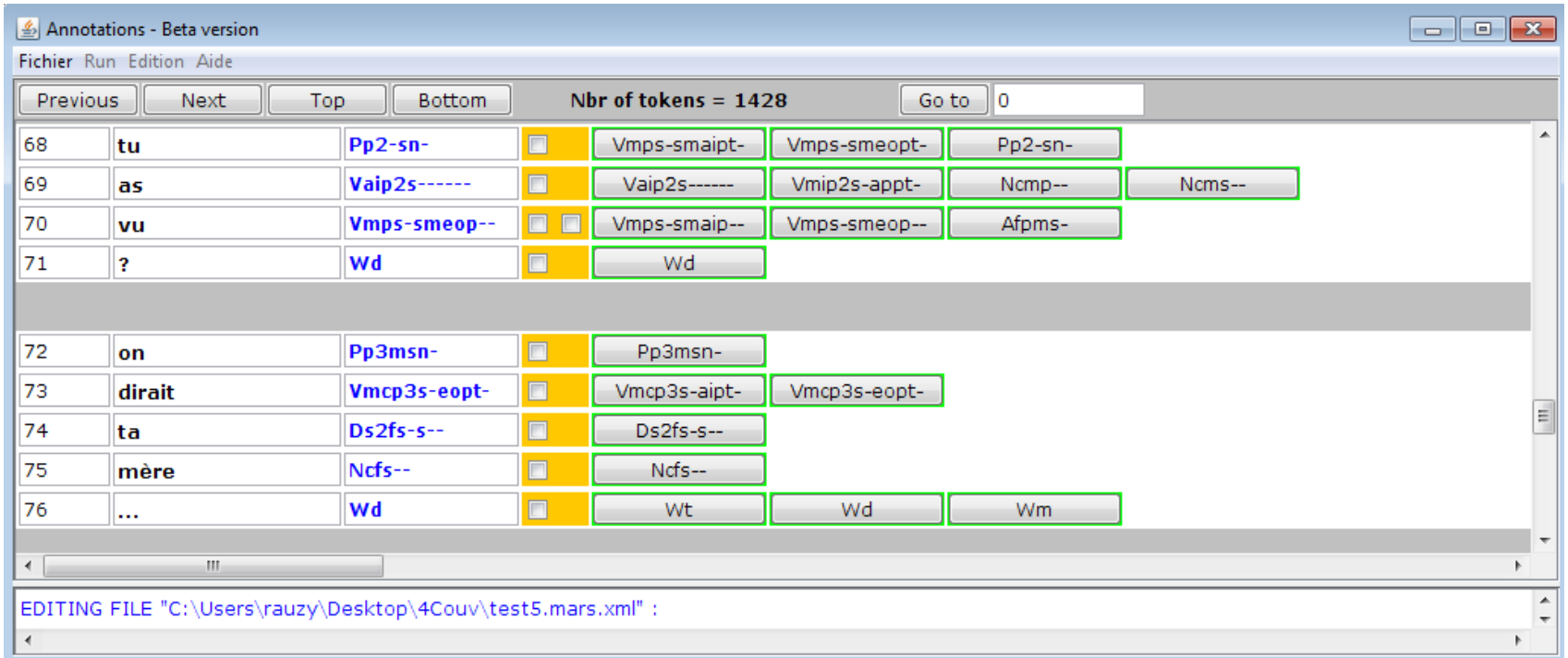
[01] Le cycle d'Elric

mot	contexte	correction	POS	morpho
fantasy	Michael Moorcock a donné vie au personnage le plus emblématique de la **fantasy** post-Tolkien : Elric , incarnation du Champion éternel , idéaliste et libertaire , plongé dans la guerre sans merci entre l' Ordre et le Chaos .	fantasy	A-	Ak-fp-
		comment :		
post-Tolkien	Michael Moorcock a donné vie au personnage le plus emblématique de la fantasy **post-Tolkien** : Elric , incarnation du Champion éternel , idéaliste et libertaire , plongé dans la guerre sans merci entre l' Ordre et le Chaos .	post-Tolkien	Nc	Ncfp--
		comment :		

[04] Le monde, tous droits réservés

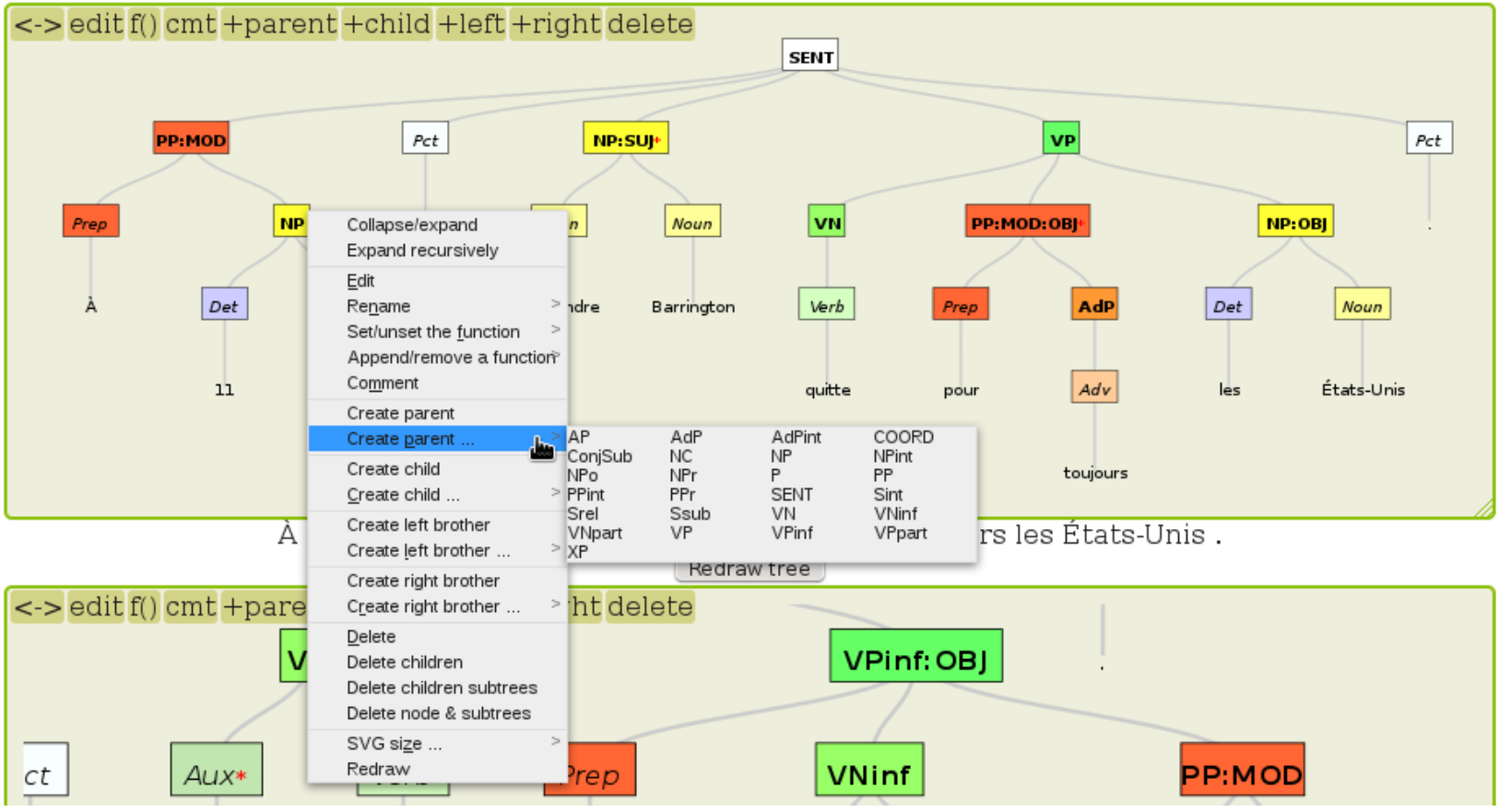
mot	contexte	correction	POS	morpho
copyrighter	Imaginez un monde ou les organes de presse auraient le pouvoir de **copyrighter** l' information ...	copyrighter	Vmn-	Vmn-----
		comment :		

Morphosyntactic tags



- one token per line
- a button for each alternative tag
- the field is also editable

Constituent trees editor

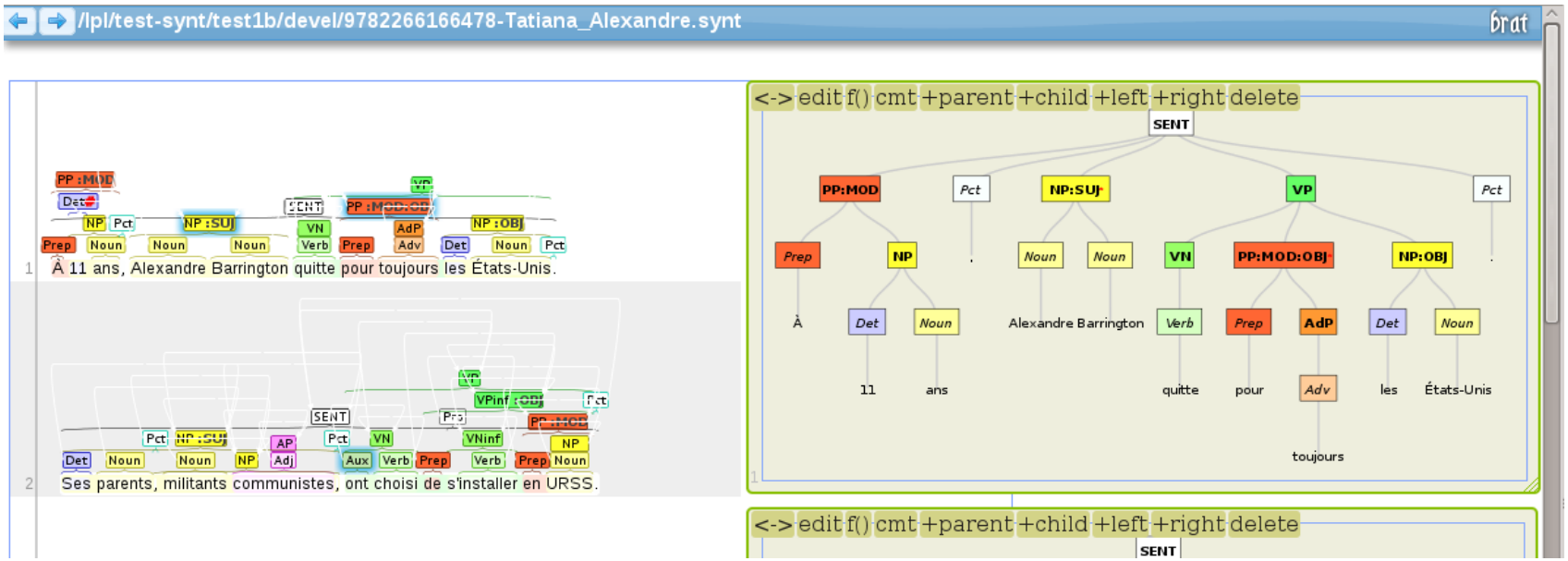


- SVG : resizable, zoomable
- drag&drop to move sub-trees
- create/delete/edit nodes
- action on various nodes

Trees editor library

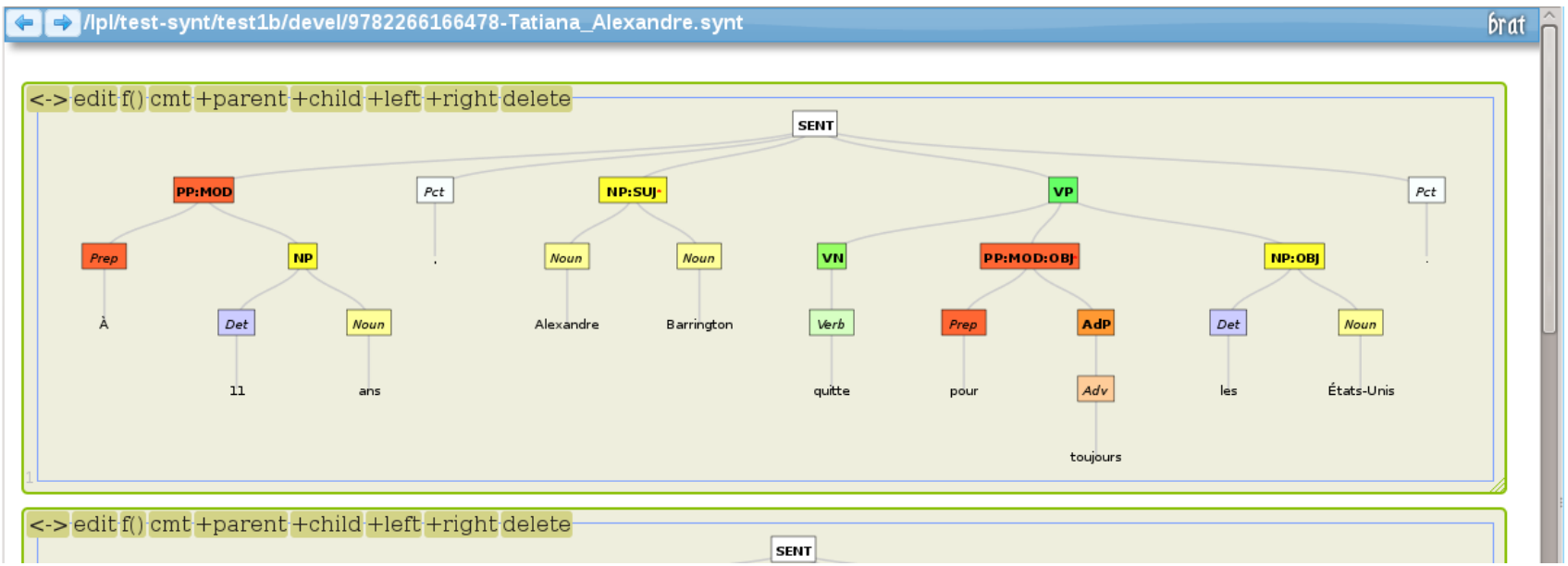
- Javascript, using open source 3rd-part libraries : *d3.js*, *jQuery*
- **standalone** : in a single HTML page, without server
- **embeded in other annotation plateforms** :

brat, WebAnno *(work in progress)*



Trees editor library

- Javascript, using open source 3rd-part libraries : *d3.js*, *jQuery*
- **standalone** : in a single HTML page, without server
- **embeded in other annotation plateforms** :
  brat, WebAnno *(work in progress)*



Perspectives

- Eye-tracking experiments
- Discursive structure studies (Prévot et al. (2015))

References I

Abeillé, A., L. Clément, and F. Toussenel. 2003. "Building a treebank for French." In *Treebanks*, ed. A. Abeillé. Kluwer, Dordrecht.

Demberg, Vera, and Frank Keller. 2008. "Data from eye-tracking corpora as evidence for theories of syntactic processing complexity." *Cognition* 109 (2): 193–210.

Paroubek, P., and M. Rajman. 2000. "MULTITAG, une ressource linguistique produit du paradigme d'évaluation." In *Actes de Traitement Automatique des Langues Naturelles*, 297–306. Lausanne, Suisse.

Prévot, Laurent, Anaïg Pénault, Grégoire Montcheuil, Stéphane Rauzy, and Philippe Blache. 2015. "Discourse Structure of Backcovers: A pilot study." In *First TextLink Action Conference*. Louvain-la-Neuve, Belgium.

References II

Rabiner, L. R. 1989. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition." *Proceedings of the IEEE* 77: 257–286.

Rauzy, Stéphane, and Philippe Blache. 2012. "Robustness and processing difficulty models. A pilot study for eye-tracking data on the French Treebank." In *Proceedings of Workshop on Eye-tracking and Natural Language Processing at The 24th International Conference on Computational Linguistics (COLING)*.

Schluter, Natalie, and Josef van Genabith. 2007. "Preparing, restructuring, and augmenting a french treebank: Lexicalised parsers or coherent treebanks?" In *Proceedings of PACLING 07*, 200–209.