

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



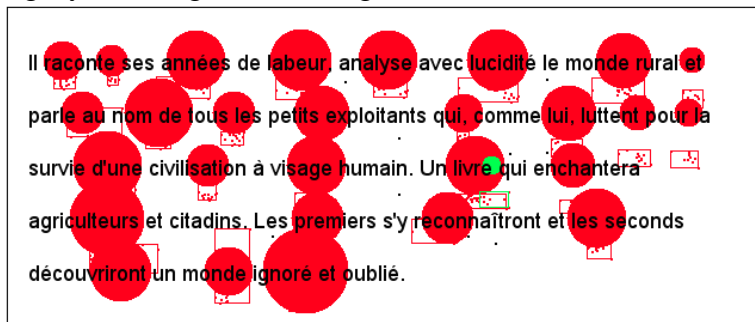
Outline

- ▶ 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

⇒ text from *backcovers*

Context

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

⇒ compatibility with the French Treebank (*FTB*, ~ 20.000 trees; Abeillé, Clément, and Toussnel (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*)

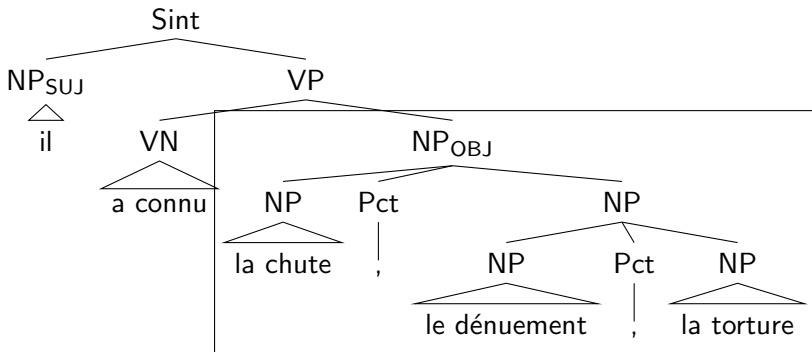
Syntactic annotation

- ▶ 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 (\neq FTB)

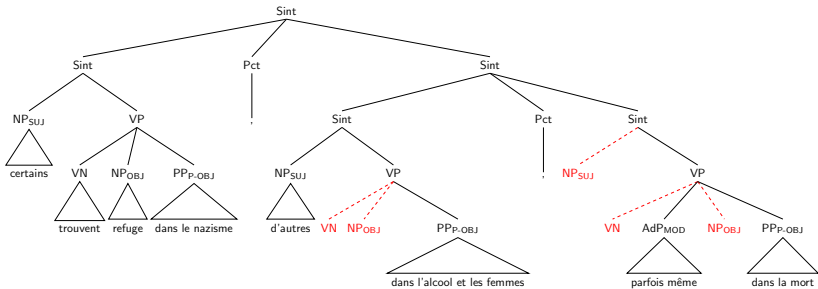
“il a connu la chute, le dénuement, la torture”

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



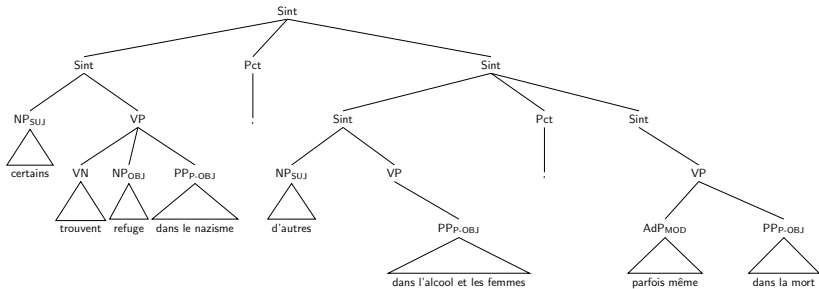
No empty category

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



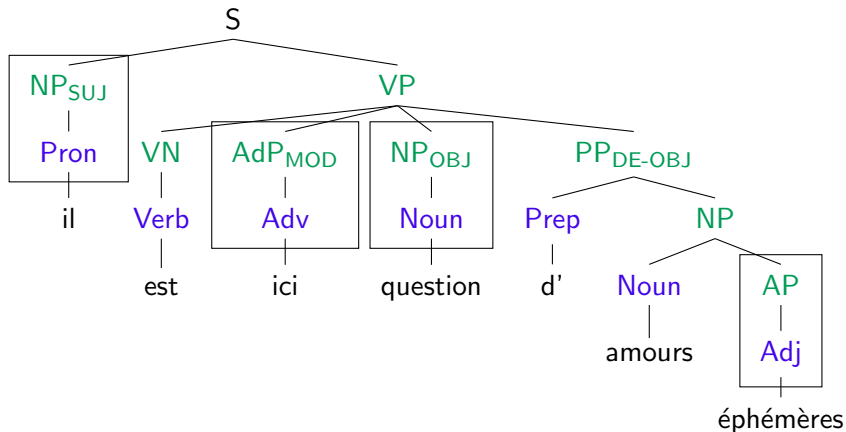
No empty category

- 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)



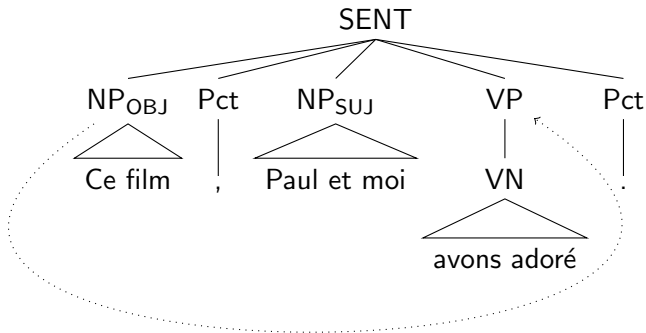
Lexical & syntactic levels

- ▶ Distinct **lexical** & **syntactic** level \Rightarrow 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
- ▶ ~ 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)): ~ 700,000 tokens
- ▶ Parser train on a LPL version of the *MFT*: 1,500 sentences

Trebanking 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

The screenshot shows the '4Couv selector' TiddlyWiki interface. The main content area displays the entry for '[01] Mother' by Luc Lang. The interface includes a sidebar with navigation options and a main content area with sections for 'Informations', 'Description', and 'Phrases'.

4Couv selector Un TiddlyWiki pour sélectionner les quatrièmes de couvertures → livres autres aide save /jai fini :-)

Pour commencer x [01] Mother x close close others edit more

[01] Mother
Last modification by 4Couv, Wednesday 03 September 2014 at 14:00:00 tags: 4Couv

Informations

| | | | | |
|-------------------------------|------------------------------|--------------------------------|-------------------------|--------------------------------|
| Auteur: Luc Lang | Titre: Mother | Tome: | | |
| Série: Roman | Sous-titre: | Langue originale: | Traduction: | |
| ISBN: 9782070454594 | Éditeur: Gallimard | Collection(s): Folio | Format: poche | Parution: 2014-08-28 |

Description
[rien](#) [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.

Phrases
[edit](#) [aide](#)

(1) «... 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.

(2) Sa mère qui se recompose comme l'horizon de sa finitude et de son destin.

16/25

Rapid evaluation

Form with boxes and list of choices

Evaluation

directives

Intérêt du texte :

0 1 2 3 4 5 6 7 8 9 10

Complexité syntaxique :

0 1 2 3 4 5 6 7 8 9 10

Difficulté discursive :

0 1 2 3 4 5 6 7 8 9 10

Note globale :

0 1 2 3 4 5 6 7 8 9 10

Genre de 4ème de couverture:

extrait+commentaire

Commentaires:

save changes

- ?
- commentaire
- genèse
- résumé
- début
- extrait
- genèse+commentaire
- résumé+commentaire
- début+commentaire
- extrait+commentaire

Sentence split and sections

Phrases

ed: 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

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

Luc Lang compose avec Mother un chant d'amour-haine surtout déclaration d'amour au père choisi.

Avec une écriture d'une rare acuité et un humour cinglant son sujet.

Pour commencer x [01] Mother x book9782070454594-sentences |

done cancel delete

book9782070454594-sentences

book9782070454594-sentences

```
[«... 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.]
```

sentences

Type tags separated with spaces, [[use double square brackets]] if necessary, or add existing tags

Wiki syntax :

- ▶ sentences are table rows
- ▶ sections separated by blank lines

Review unknown words

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

close close others edit more

Mots inconnus

Last modification by [4Couv](#), Wednesday 03 September 2014 at 14:00:00

[01] Le cycle d'Elric

tags:
MenuMore

| 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 <input type="checkbox"/> | A- <input type="checkbox"/> | Ak-fp- <input type="checkbox"/> |
| comment : <input type="text"/> | | | | |
| 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 <input type="checkbox"/> | Nc <input type="checkbox"/> | Ncfp-- <input type="checkbox"/> |
| comment : <input type="text"/> | | | | |

[04] Le monde, tous droits réservés

| mot | contexte | correction | POS | morpho |
|-----------------------------------|---|---|----------------------------------|--------------------------------------|
| copyrighter | Imaginez un monde où les organes de presse auraient le pouvoir de copyrighter l' information ... | copyrighter <input type="checkbox"/> | Vmn- <input type="checkbox"/> | Vmn----- <input type="checkbox"/> |
| comment : <input type="text"/> | | | | |

Morphosyntactic tags

The screenshot shows the 'Annotations - Beta version' window. At the top, there are menu options: 'Fichier', 'Run', 'Edition', 'Aide'. Below the menu is a toolbar with 'Previous', 'Next', 'Top', and 'Bottom' buttons. The status bar indicates 'Nbr of tokens = 1428' and 'Go to 0'. The main area displays a list of tokens with their corresponding morphosyntactic tags and alternative tag buttons.

| Line | Token | Tag | Buttons |
|------|--------|--------------|--|
| 68 | tu | Pp2-sn- | Vmps-smaipt- Vmps-smeopt- Pp2-sn- |
| 69 | as | Vaip2s----- | Vaip2s----- Vmip2s-appt- Ncmp-- Ncms-- |
| 70 | vu | Vmps-smeop-- | Vmps-smaip-- Vmps-smeop-- Afpms- |
| 71 | ? | Wd | Wd |
| | | | |
| 72 | on | Pp3msn- | Pp3msn- |
| 73 | dirait | Vmcp3s-eopt- | Vmcp3s-aipt- Vmcp3s-eopt- |
| 74 | ta | Ds2fs-s-- | Ds2fs-s-- |
| 75 | mère | Ncfs-- | Ncfs-- |
| 76 | ... | Wd | Wt Wd Wm |

At the bottom, the status bar shows 'EDITING FILE "C:\Users\rauzy\Desktop\4Couv\test5.mars.xml" :

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

Constituent trees editor

The screenshot displays a constituent tree editor interface. At the top, a command bar contains: `<-> edit f() cmt +parent +child +left +right delete`. The main area shows a tree structure for the sentence "A... Barrington... quitte pour toujours les États-Unis.". The root node is **SENT**. Its children are **PP:MOD**, **Pct**, **NP:SUJ**, **VP**, and **Pct**. The **NP:SUJ** node has children **NP** (with children **Prep** 'A' and **Det** '11') and **Noun** 'Barrington'. The **VP** node has children **VN** 'quitte', **PP:MOD:OBJ** (with children **Prep** 'pour' and **Adv** 'toujours'), and **NP:OBJ** (with children **Det** 'les' and **Noun** 'États-Unis'). A context menu is open over the **NP** node, listing various actions like 'Collapse/expand', 'Create parent...', and 'Create child...'. A sub-menu is also visible, listing node types such as **AP**, **ConjSub**, **NPo**, **PPint**, **Srel**, **VNpart**, **XP**, **AdP**, **NC**, **NPr**, **P**, **PPr**, **Seub**, **VP**, **AdPint**, **NP**, **NPint**, **SENT**, **VN**, **VPint**, **COORD**, **NPint**, **PP**, **Sint**, **VNinf**, and **VPpart**. A 'Redraw tree' button is at the bottom of the menu.

- ▶ 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
- ▶ **embedded in other annotation platforms** :



brat,



WebAnno

WebAnno (*work in progress*)

1 À 11 ans, Alexandre Barrington quitte pour toujours les États-Unis.

2 Ses parents, militants communistes, ont choisi de s'installer en URSS.

Syntax tree for sentence 1:

```
graph TD
    SENT1[SENT] --- PP_MOD[PP-MOD]
    SENT1 --- Pct1[Pct]
    SENT1 --- NP1[NP]
    SENT1 --- VP[VP]
    SENT1 --- Pct2[Pct]
    PP_MOD --- Prep[Prep]
    NP1 --- Det[Det]
    NP1 --- Noun1[Noun]
    NP1 --- Noun2[Noun]
    NP1 --- Noun3[Noun]
    VP --- VN[VN]
    VP --- PP_MOD_OBJ[PP-MOD:OBJ]
    VP --- NP_OBJ[NP:OBJ]
    Prep --- À[À]
    Det --- 11[11]
    Noun1 --- ans[ans]
    Noun2 --- Alexandre[Alexandre]
    Noun3 --- Barrington[Barrington]
    VN --- quitte[quitte]
    PP_MOD_OBJ --- Prep2[Prep]
    PP_MOD_OBJ --- Adv[Adv]
    NP_OBJ --- Det2[Det]
    NP_OBJ --- Noun4[Noun]
    Prep2 --- pour[pour]
    Adv --- toujours[toujours]
    Det2 --- les[les]
    Noun4 --- États-Unis[États-Unis]
```


Trees editor library

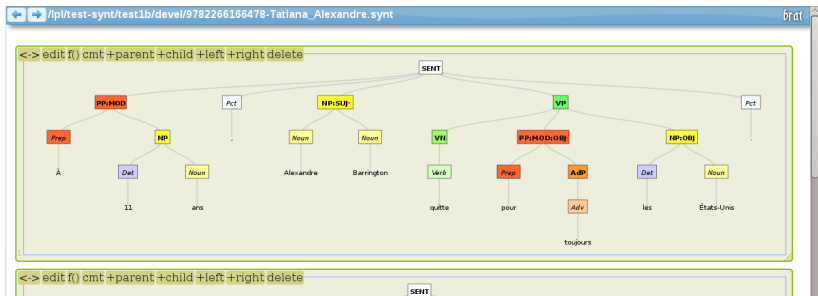
- ▶ Javascript, using open source 3rd-part libraries : *d3.js*, *jQuery*
- ▶ **standalone** : in a single HTML page, without server
- ▶ **embedded in other annotation platforms** :



brat,



WebAnno (*work in progress*)



Perspectives

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

References I

Abeillé, A., L. Clément, and F. Toussanel. 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.

Prévoit, 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.