

Corpus arborés et parsing

Cours 10

Santiago Herrera
s.herrera@parisnanterre.fr

Ce que nous avons fait jusqu'à présent

Nous partons du langage naturel

du coup je parle une histoire
parce qu'en fait, pour moi, le truc qui fait une génération, c'est vraiment, genre, euh.

les faits !!?? come le fait que sur les 19 pirates de l air, certains se portent encore comme des charmes dans leur pays ? il s agit clairement de la version officielle, des les premieres lignes de l article. il serait bon de faire effectivement un article avec que les faits incontestes (2 avions dans l pirate) figurez_vous qu'il y a, en général et sur wikipedia en particulier, quantité de faits qui n officiel pas justifiés par une publication peer-review. Peut-être que le fait de publier comme ça me Levochik4 sentir les limites d'application de cette exigence ? Je suis bien évidemment sérieux quand demande sur la BBC. Et votre raccourci qui consiste à transformer "une recherche de 30s suff trouver une source pertinente" en "j'ai fait des recherches de 30s." est intellectuellement malhonnête. Levochik4 novembre 2009 à 11:33 (CET)

Nous partons du langage naturel

du coup je parle une histoire

parce qu'en fait, pour moi, le truc qui fait une génération, c'est vraiment, genre, euh.

les faits !!?? come le fait que sur les 19 pirates de l air, certains se portent encore comme des charmes dans leur pays ? il s agit clairement de la version officielle, des les premieres lignes de l article. il serait bon de faire effectivement un article avec que les faits incontestes (2 avions dans les tours, qui s ecroulent, un truc au Pentagone etc.), une version officielle (al qaeda, 19 pirates, pas de wtc 7, avion sur le pentagone etc.) et un article " contestation de la version officielle", qui irait des theories dites du complot aux simples pointages d'invraisemblances.
Levochik27 septembre 2009 à 10:41 (CEST)



ou de chaînes de caractères

Unités d'analyse

- Segmentation en phrases
- Ensuite... Mots ? Morphèmes ? Tokens !

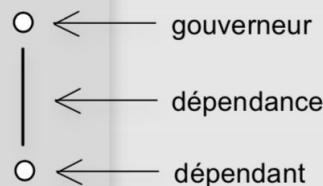
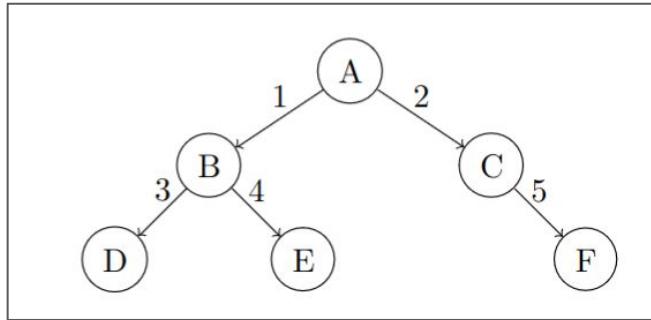
Bonne idée! à propos je propose de faire la même chose pour Naufrage du Titanic, après tout il existe aussi des théories qui disent qu'il n'a jamais coulé, et puis on pourra faire pareil pour Apollo 11, une quantité importante de personnes affirme que tout est faut. Pour chaque fait historique il existe ce genre de chose, il y a toujours des personnes pour nier l'évidence. Pourquoi pas changé l'article sur la Shoah aussi... Wikipédia aspire à devenir un encyclopédie de référence, et cette discussion semble l'en écarter. D'ailleurs je trouve que l'article prend trop partie en faveur des conspirasnistes, comme je l'ai dis plus bas. Guillaume70 Guillaume 30 septembre 2009 à 10:43 (CEST)

Analyse d'une phrase

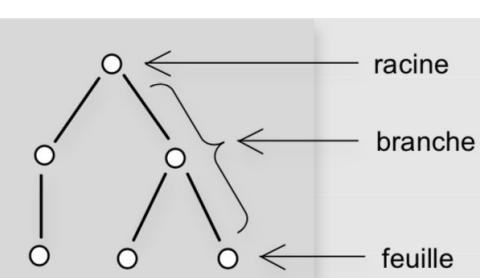
À propos, je propose de faire la même chose pour Naufrage du Titanic

- Des éléments qui dépendent les uns des autres
- Des mots plus importants que d'autres avec distributions différentes.
- Une hiérarchie et un ordre
- etc.

Structure : des arbres ?

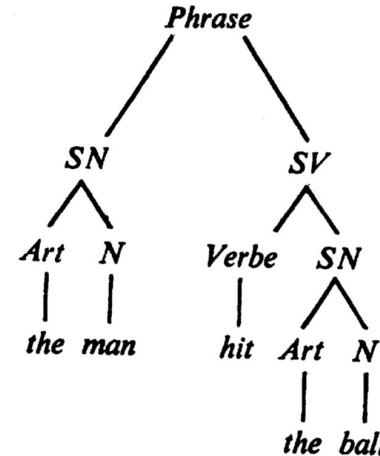
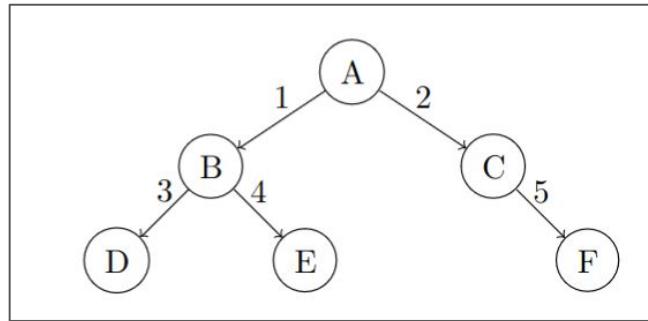


Dépendance

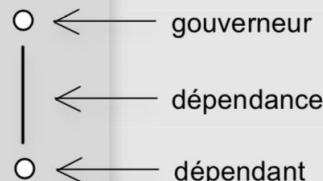
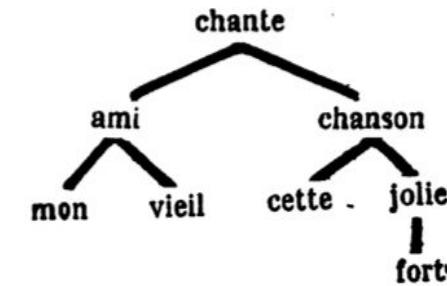


Branche

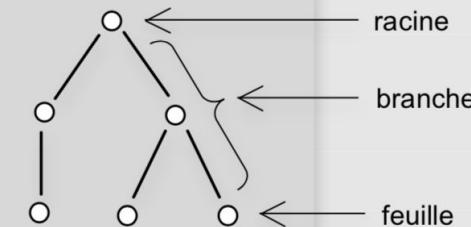
Structure : des arbres !



Arbre de constituants
Chomsky 1957 (version 1969)



Dépendance



Branche

Arbre de dépendance
Tesnière 1959

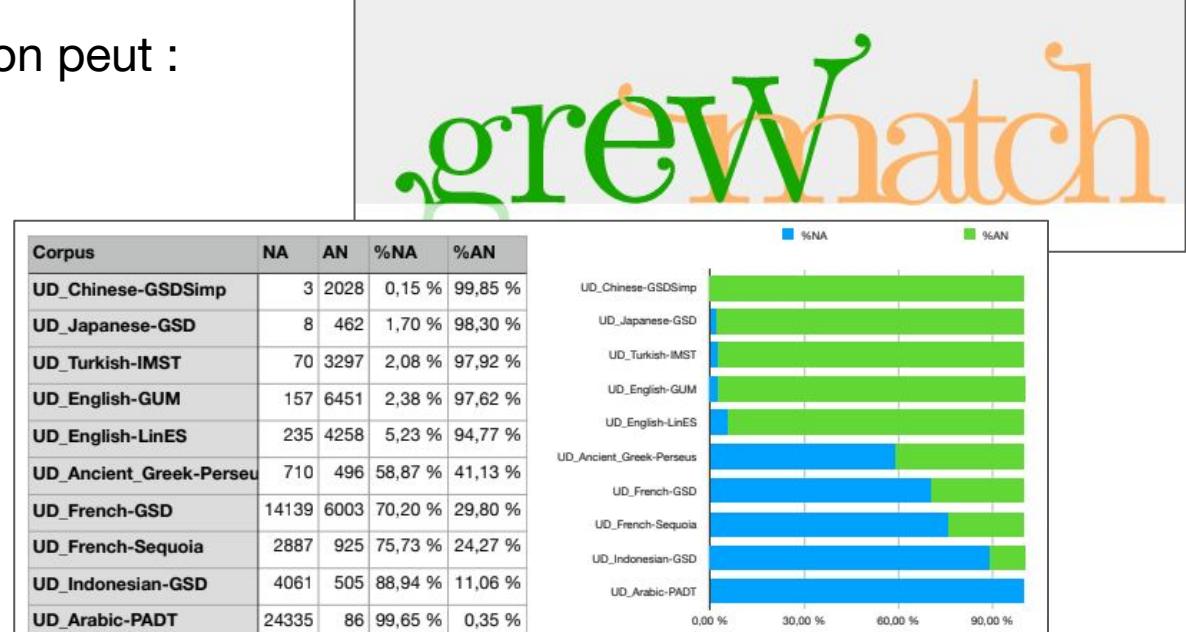
Mais...

Quel est l'objectif de la représentation de la langue dans les arbres ?

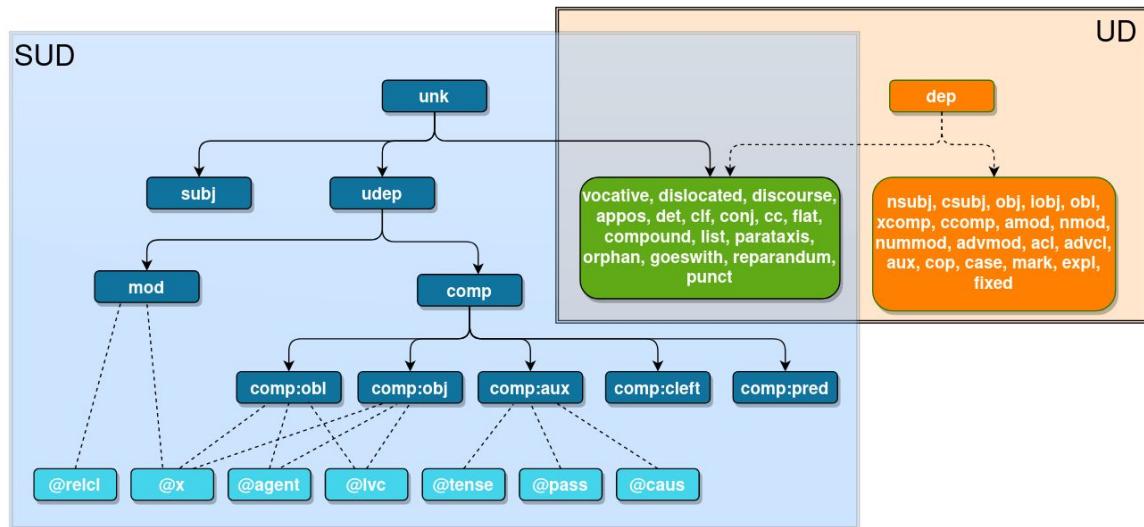
- **Nous modélisons ce que nous pensons être la structure de la langue.**

Mais, en plus, maintenant, on peut :

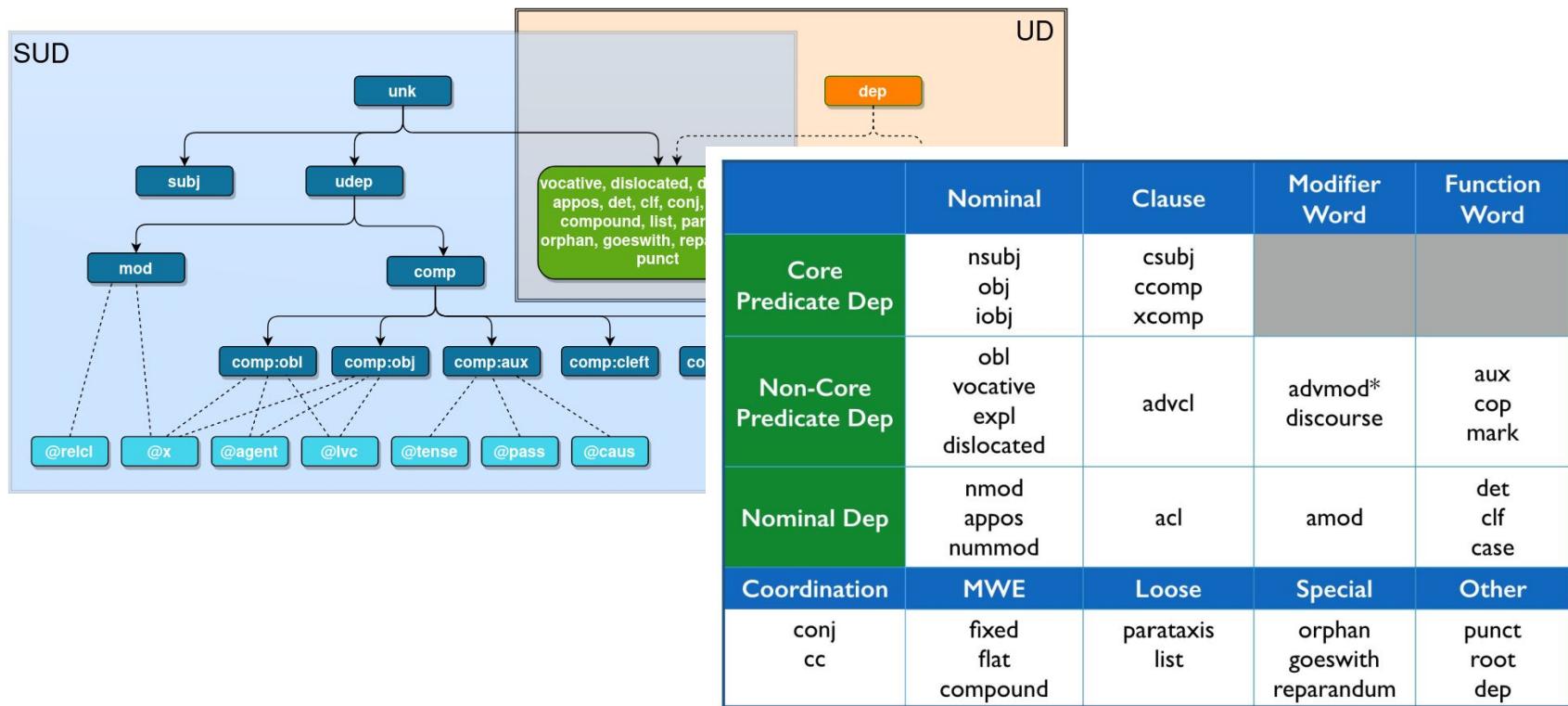
- **Faire de calculs**
- **Formaliser le langage**
- **L'informatiser**
- **Faire des requêtes**



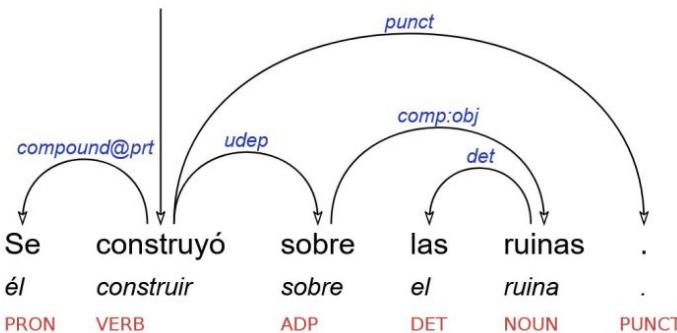
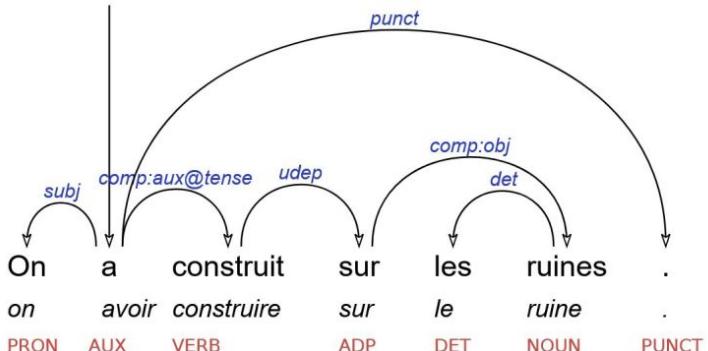
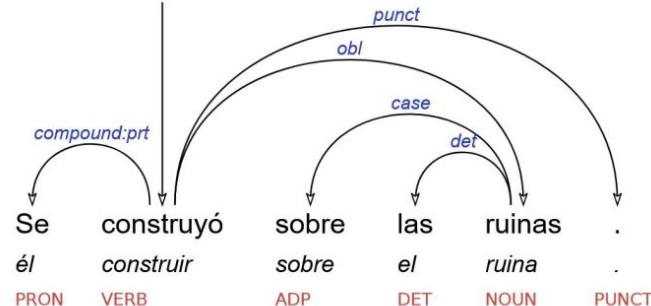
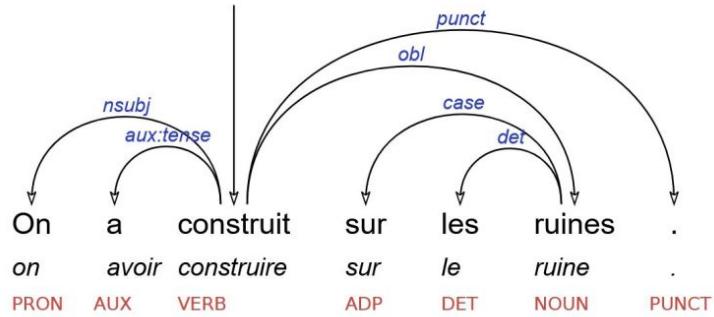
Modéliser == Catégoriser



Modéliser == Catégoriser

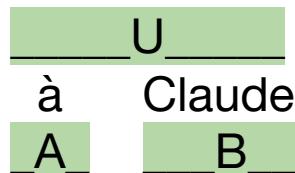
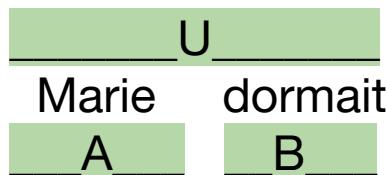


Modéliser == Hiérarchiser

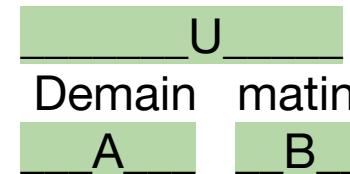


Trouvez les bons critères

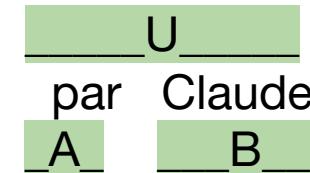
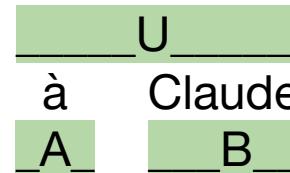
Critère distributionnel avec effacement négatif



Critère distributionnel avec effacement positif

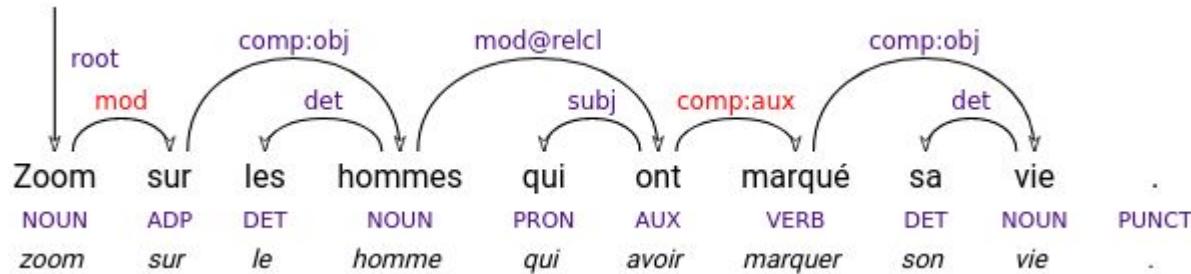


Critère distributionnel sans effacement



Annotation

- Mettre en pratique l'ensemble du dispositif théorique et voir comment il s'ajuste à la langue.



- Nous n'avons pas évalué l'accord entre annotateurs...
(vous verrez ça dans d'autre cours)

Tout est parfait

On simplifie les phénomènes linguistiques et discursifs

- Seulement 3 niveaux d'analyse
- Certaines informations sont laissées de côté

Toujours dans un compromis entre la complexité du modèle et la complexité du langage

Enfin, oui...

- On peut l'encoder

```
# global.columns = ID FORM LEMMA UPOS XPOS FEATS HEAD DEPREL DEPS MISC
# sent_id = fr-ud-train_06412
# text = Je le conseille vivement.
1 Je moi PRON _ Emph=No|Number=Sing|Person=1|PronType=Prs 3 subj _ wordform=je
2 le lui PRON _ Emph=No|Gender=Masc|Number=Sing|Person=3|PronType=Prs 3 comp:obj
3 conseille conseiller VERB _ Mood=Ind|Number=Sing|Person=1|Tense=Pres|VerbForm=Fin 0 root -
4 vivement vivement ADV _ 3 mod _ SpaceAfter=No
5 . . PUNCT _ 3 punct _ -
```

- Pour l'exploiter il faut définir la structure de données : graphes, structures de traits, etc.
- On a appris un langage de requête et à utiliser un outil
 - Pour explorer des arbres
 - Clustériser des résultats
 - Pour modifier les arbres
- On a exploré des arbres avec Python

Mais... maintenant ?

Tout ça pour quoi ?

- Pour comprendre le fonctionnement de la langue
- Disposant d'un modèle de syntaxe motivé linguistiquement, des phrases structurée et annotée, sur lesquelles on peut faire des opérations et de calculs...
- On peut donc faire des prédictions...

Tout ça pour quoi ?

- Pour comprendre le fonctionnement de la langue
- Disposant d'un modèle de syntaxe motivé linguistiquement, des phrases structurée et annotée, sur lesquelles on peut faire des opérations et de calculs...
- On peut donc faire des prédictions...

PARSING

Pourquoi le parsing automatique ?

Automatiser l'analyse syntaxique permet d'obtenir rapidement la structure syntaxique de grands corpus.

Pour :

- Des **outils** qui ont besoin d'une analyse fine de la phrase pour fonctionner|
(ex : génération de brevets)
- Des **linguistes** qui étudient les constructions d'une langue
- Des **apprenants** qui veulent comprendre en détail une phrase
- Des **professeurs de langues** qui construisent leurs cours en requérant des corpus
- ~~- Des modèles de traduction qui utilisent la structure syntaxique pour produire la traduction~~