

Corpus arborés et parsing

M1 - pluriTAL

Aleksandra Miletic - Chercheuse CNRS
Ioana Madalina Silai - Doctorante

Cours

- 12 séances
 - 6 séances avec Madalina
 - 6 séances avec Aleksandra
- Besoin d'un ordinateur
- Concepts théoriques et exercices pratiques
- Matériel partagé dans un Drive:

<https://drive.google.com/drive/folders/1OZd2TTntcQUI9hYwKWzrw9ibD6LWECm4?usp=sharing>

Évaluation

2 DST

- 27 Octobre
- 19 *Décembre*

Contact

imsilai@parisnanterre.fr
amiletic@parisnanterre.fr



Mon parcours

- Licence Français et Allemand à l'Université de York (Royaume Uni)
- PGCE (Master de Pédagogie) + Enseignement des langues au lycée
- Master TAL
- Doctorat - La catégorisation grammaticale: caractérisation et induction (directeur: Sylvain Kahane)

Votre tour

- **Linguistique ?**
- **TAL ?**
- **Langues vivantes ?**

Introduction

De quoi parlons-nous lorsque nous parlons de ...

Corpus arborés et parsing ?

De quoi parlons-nous lorsque nous parlons de ...

Corpus arborés et parsing ?

Corpus



- *Corpus* (lat.) : ‘corps’ ; ‘collection’
- Un ensemble de textes ou discours produits
 - => données linguistiques attestées
- Divers types de corpus en fonction du contenu, de la finalité, etc.
 - ex. Corpus de l’oral
 - ex. Corpus parallèles

et on est, on était six dans le maison.
enfin c'est pas, c'est pas dans la maison, c'est euh
il y a une maison, et une cour.

Chapter 006, Sir Jonathan Sacks

The **syntax** is fractured .

es

La sintaxis está cortada .

fr

La grammaire n' est pas correcte , la syntaxe est fracturée .

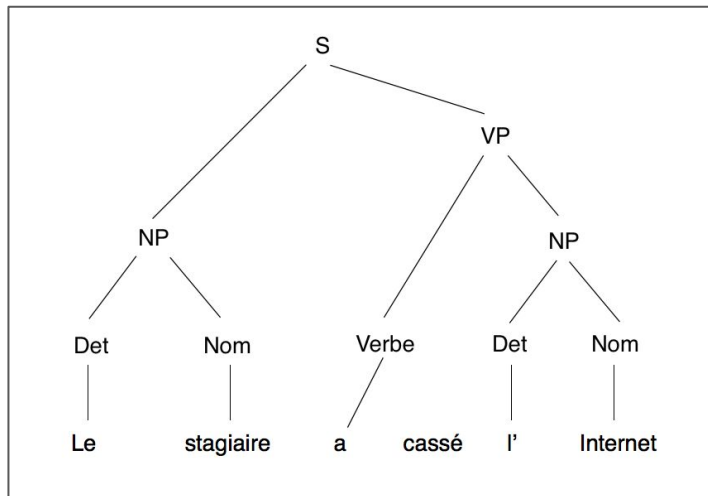
- Un bon corpus doit comporter des métadonnées
- Jeu de données (*dataset*) vs Corpus
- Format numérique

**Connaissez vous d'autres
types de corpus?**

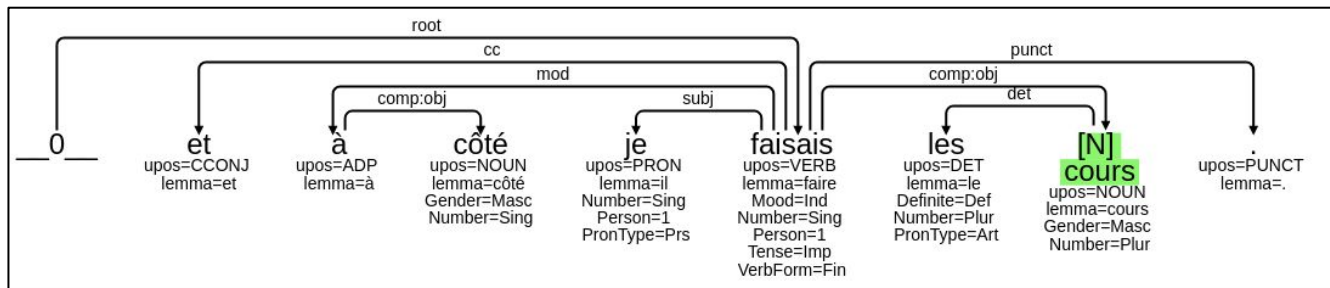
De quoi parlons-nous lorsque nous parlons de ...

Corpus arborés et parsing ?

Corpus arborés (en syntaxe) / treebanks






- Des phrases associées à des arbres syntaxiques
 - Phrase typiquement **étiquetée** (en plusieurs types d'information)
 - La forme des arbres dépend du **cadre théorique** adoptée
 - **Annotation** (ou correction) **à la main** par des experts
 - L'ensemble de décisions prises pendant l'annotation est guidé par des **choix théoriques**



De quoi parlons-nous lorsque nous parlons de ...

Corpus arborés et parsing ?

Parsing

- En **psycholinguistique** : Le parsing implique l'analyse et la compréhension d'un énoncé
- En **informatique** : Analyse d'une chaîne de symboles ou caractères (strings) pour relever sa structure. Souvent avec une grammaire (des règles)
 - e.g. N'importe quel langage de programmation.
 -  Parsez **2+3x5**
- En **syntaxe/TAL** : Analyse automatique d'une phrase (ou d'une autre unité de segmentation) afin de trouver sa structure et de catégoriser ses éléments.
 -  Parsez **I saw a woman with a telescope wrapped in paper**
 -  Parsez **J'ai vu une femme avec des jumelles**

De quoi parlons-nous lorsque nous parlons de ...



Corpus arborés et parsing ?

The diagram consists of two rectangular boxes. The left box contains the text 'Corpus arborés' and the right box contains the text 'parsing ?'. A curved arrow originates from the top of the right box and points to the top of the left box, indicating a relationship or flow from parsing back to the corpus.

Création des treebanks

- Un parseur s'entraîne avec des treebanks

Programme de départ

- Introduction aux concepts fondamentaux
- Création et exploration d'un treebank et requêtage
- Annotation syntaxique
- Création d'un schéma d'annotation et l'accord inter-annotateur
- Explorations statistiques d'un treebank (avec python)
- Parsing et bootstrapping



Histoire

Diagrammes syntaxiques

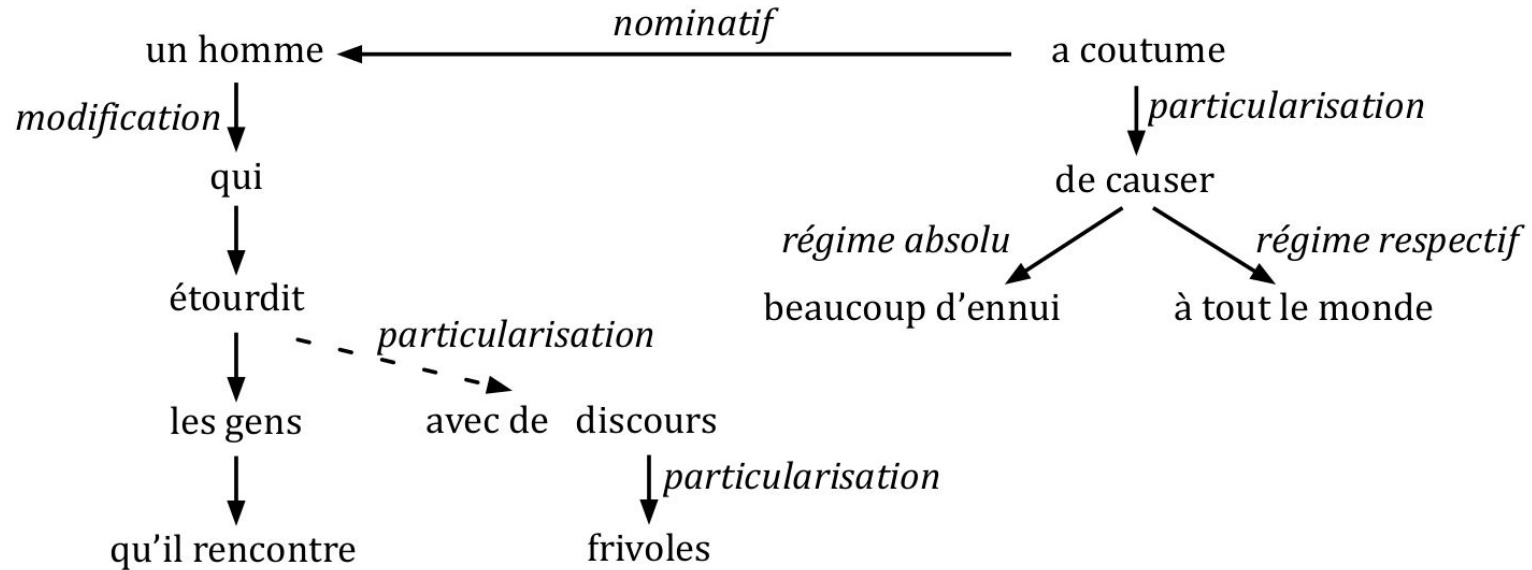
Claude Buffier (1709)

Un homme qui étourdit les gens qu'il rencontre avec de frivoles discours, a coutume de causer beaucoup d'ennui à tout le monde. Je dis que dans ce discours, tous les mots sont pour modifier le nom **un homme**, & le verbe **a coutume**, & que c'est en cela que consiste tout le mystère & toute l'essence de la syntaxe des langues :

- 1° le nom **un homme**, est modifié d'abord par le **qui déterminatif** : car il ne s'agit pas ici d'un homme en général, mais d'**un homme marqué & déterminé** en particulier par l'action qu'il fait d'**étourdir** ;
- de même il ne s'agit pas d'un homme **qui étourdit** en général, mais **qui étourdit** en particulier les gens, & non pas **les gens** en général, mais en particulier **les gens qu'il rencontre**.
- Or cet homme qui étourdit ceux qu'il rencontre, est encore *particularisé* par **avec des discours**, & **discours** est encore *particularisé* par **frivoles**.
- On peut voir le même dans la suite de la phrase : **a coutume** est *particularisé* par **de causer**, **de causer** est *particularisé* par ses deux *régimes*, par son *régime absolu*, savoir, **beaucoup d'ennui**, & par son *régime respectif*, **à tout le monde**.

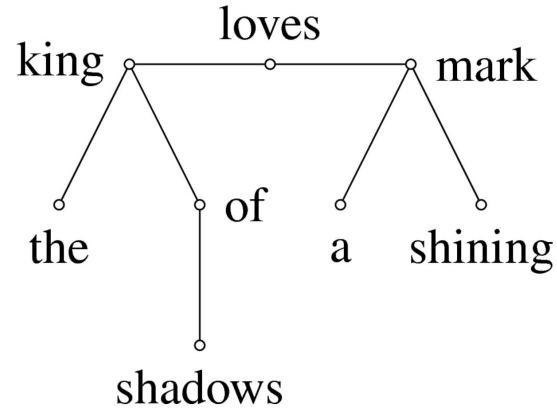
Voilà donc comment tous les mots d'une phrase quelque longue qu'elle soit, ne sont que pour modifier le nom & le verbe.

Diagrammes syntaxiques : Claude Buffier (1709)



Diagrammes syntaxiques : Stephen W. Clark (1847)

1. “*The king of shadows loves a shining mark.*”
(13.)



Une structure de dépendance dé-réifiée

Diagrammes tabulaires

Reproduction d'un tableau d'analyse grammaticale par **Louis Gaultier** (1817, 11) : Le Père et la Mère de Zoé sortirent un matin, lorsque le Soleil commençait à paraître sur l'Horizon, pour aller voir un de leurs amis qui avait été indisposé.

Exemple de PHRASES décomposées ?
dans le TABLEAU d'Analyse de Grammaire, d'après la Méthode de L. GAULTIER.

MOTS de la PHRASE à ANALYSER.	DIVISION générale des MOTS.		Rapports généraux du Sujet			Rapports généraux du Verbe Simple.				DIVISION des Mots-Parties de la PHRASE.	MEMBRES de la PHRASE analysée.
	1. Préfixe de mot?	2. Préfixe de phrase?	3. Préfixe de genre?	4. Préfixe de nombre?	5. Préfixe de cas?	6. Préfixe de mode?	7. Préfixe de temps?	8. Préfixe de lieu?	9. Préfixe de fin?		
Le	P.	P.								Article Simple.	que?
Père	N.	S.	M.	S.	M.		(de mot.)			Commun	
et	P.	C.								Conjonctive Simple.	
la	P.	P.								Article Simple.	
Mère	N.	S.	F.	S.	M.		(de mot.)			Commun	
de	P.	P.								P. S. Simple.	que firent-ils?
Zoé	N.	S.	F.	S.	G.		(Dépendant du Substantif Mère.)			Propre	
sortirent	V.	S.				P.	3 P.	P ¹	ind.	Temps simple Passé défini Présent, V. Act.	
un	N.	Adj. indéf.	M.	S.	Déterminé ¹		(Modification de mot.)			Nominal Cardinal	
matin.	N.	S.	M.	S.	Déterminé ¹		(Dépend de la prép. pour avec entente.)			Commun	
lorsque	P.	C.								De Temps	pourquoi?
le	P.	P.								Article Simple.	
Soleil	N.	S.	M.	S.	M.		(de mot.)			Commun	
commençait	V.	S.				S.	3 P.	P ¹	ind.	Temps simple Imparfait 1 ^{re} conjugaison, V. B.	
à	P.	P.								Préposition Simple	
paraître	V.	J.								P ¹ à l'infinitif, V. B.	pourquoi?
sur	P.	P.								Préposition Simple	
l'	P.	P.								Article Simple.	
Horizon	N.	S.	M.	S.	Déterminé ¹		(Dépend de la prép. sur.)			Commun	
pour	P.	P.								P. S. Simple	
aller	V.	J.								P ¹ à l'infinitif, V. B.	pourquoi?
voir	V.	J.								P ¹ à l'infinitif, V. B.	
un	N.	Adj. indéf.	M.	S.	2 C.		(Modification de mot avec entente.)			Nominal Cardinal	
de	P.	P.								Préposition Simple	
leurs	N.	P.	M.	P.	G.		(Modification de mot.)			Pronom Possessif	
amis	N.	S.	M.	S.						Commun	

Diagrammes tabulaires

Analyse de phrases complexes chez Louis Gaultier (1817, 34)

CONSTRUCTION ET ANALYSE

SECTION III^e. - PHRASES COMPOSÉES.

La phrase composée est la réunion de deux phrases simples liées ensemble par un pronom relatif ou par une conjonction.

L'une s'appelle principale; l'autre s'appelle subordonnée, parce qu'elle dépend de la première.

CHAPITRE I^{er}. - PHRASE PRINCIPALE MODIFIÉE PAR UNE RELATIVE.

(N. B. Ces phrases seront caractérisées et citées par les lettres o p q.)

CONJONCTIONS Pronoms relatifs INTERJECTIONS.	(1) SUJET ET SES MODIFICATIONS.	(2) VERBE ET SES MODIFICATIONS.	(3) RÉGIME DIRECT ET SES MODIFICATIONS.	(4) RÉGIME INDIRECT ET SES MODIFICATIONS.	(5) DÉTERMINATIF ET SES MODIFICATIONS.
§ I. - Phrase principale qui précède la subordonnée relative, (o)	qui	Celui - là	est heureux		
		ne désire	rien.		
		<i>Qui? celui qui ne désire rien</i>	<i>Qu'est-ce? ses larmes.</i>		
	qui	Les bons ouvrages	seront les seuls		
		passeront		à la postérité.	
		<i>Quels? les bons ouvrages</i>	<i>Que recevront? ceux qui passeront à la postérité.</i>		
	qui	Punissez	le cruel		
		ne pardonne pas.			
		<i>Qui? Punissez? punissez</i>	<i>Qui? le cruel qui ne pardonne pas</i>		
	qui	J'	accoutume	mon âme	à souffrir ce
	ils	font.			
	<i>Qui? Je</i>	<i>Que font-ils? accoutume</i>	<i>Qu'est-ce? mon âme</i>	<i>À quoi? à souffrir ce qu'ils font</i>	
	ils	arrivent			à l'instant
	nous	quittons	cette île.		
	<i>Qui? ils</i>	<i>Que font-ils? arrivent</i>			<i>Quand? à l'instant de nous quitter cette île</i>

Diagrammes : analyse en constituants

Section sur les infinitifs en
position objet de **Otto Jespersen**
(1937, 48-49)

17. 2. Object.

He wishes to sing **S V O(I)**.

He wants to be **kind** to everybody **S V O(IPp1)**.

He is able (willing) to sing **S V P(2O(I))**.

He wants to see her **S V O(IO₂)**.

F. Il désire la voir; G. Er wünscht sie zu sehen **S V O(O₂I)**.

Ru. Dajte emu govorit' 'Give him (leave) to speak' { **SV** } **O O(I) !**

He had to go at once **S V O(I3)**.

He had to say something **S V O(IO₂)**.

F. J'ai à vous remercier **S V O(O₂I)**.

G. Sie haben zu gehorchen; Dan. De har at lystre **S V O(I)**.

It. Non avete da temere 3^a { **SV** } **O(I)**.

Many questions have to be settled **S(2*1) V O(I^b)**.

He could find it in his heart to hurt her **S V o p1(S*1) O(IO₂)**.

He promised her to go **S V O O(S*I)**.

He allowed her to go **S V O O(S₂*I)**, or, more explicitly,
S V O O(S₂*O(=O)I).

The two sentences are seemingly parallel; their different import, denoted in our symbols, naturally follows from the fact that a promise refers to one's own acts, a permission to the other person's acts.

F. Dites-lui de se hâter { **SV** } **O* O(O₂*I)**.

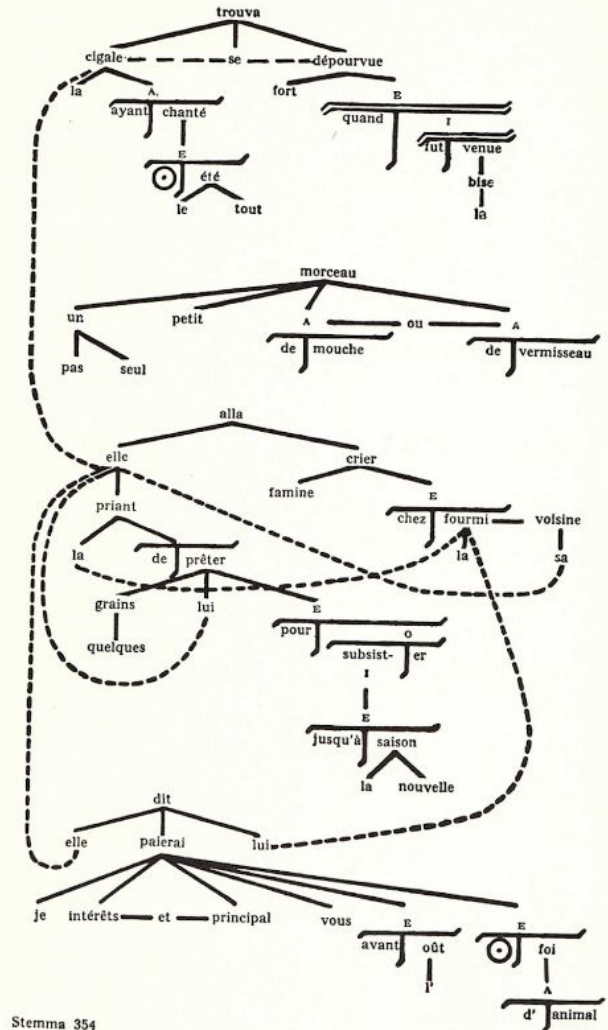
F. Il me faut aller **S O V O(IS* = O)**.

How is Sp. *que* to be symbolized in
Tengo que hablarte 'I have (something) to speak to you (about)'?

Possibly { **SV** } **O(O*IO)**.

Diagrammes en dépendance

Première moitié de l'analyse de La cigale
et la fourmi par **Lucien Tesnière** (1959 :
638)

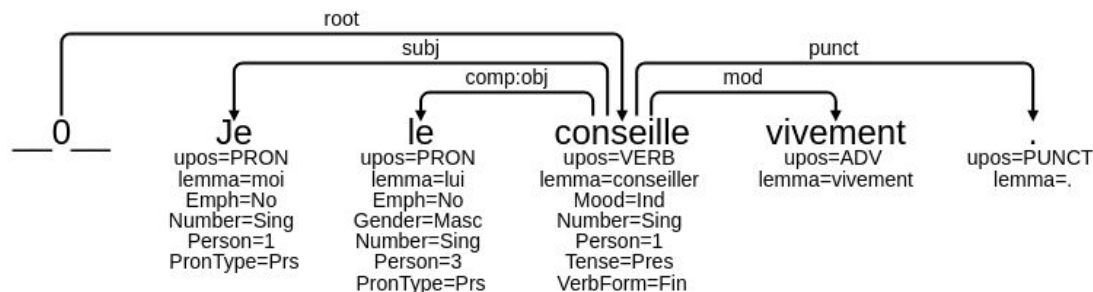


Bref historique de treebanks

- 1976: Talbanken (suédois)
- 1989-1996: **Penn** Tree Bank (anglais)
- 1997: Negra Treebank (allemand)
- 1995-now: **Prague Dependency Treebank** (tchèque)
- 2003: French Tree Bank (français, Le Monde)
- ~ 2005: L'analyse en dépendances s'impose
- 2005: Stanford parser (2002) propose une analyse en dépendances
- 2007: CoNLL dataset => format d'encodage **CoNLL** pour les arbres en dépendances
- 2008: POS interset, projets de conversion
- 2014: Google publie des treebanks pour 30 langues à partir du schéma de Stanford
- 2014: Début du projet **Universal Dependencies**

Treebanks aujourd'hui

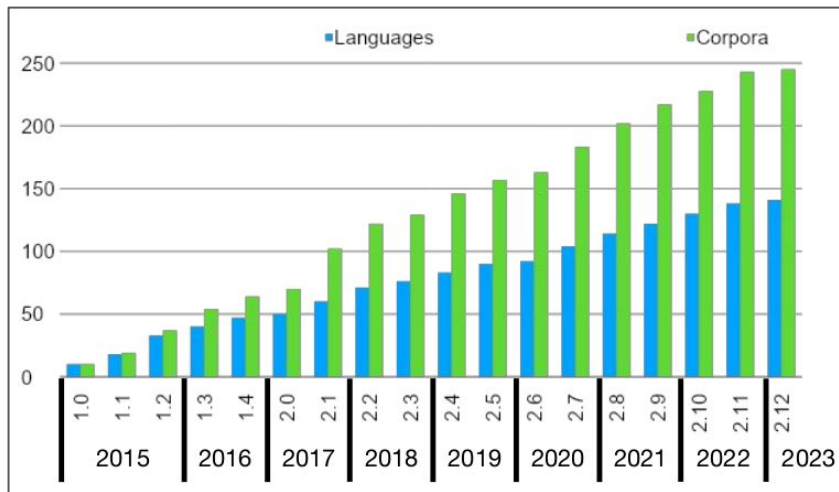
- Format numérique, requêtable et encodé dans un format de fichier **conllu**



```
# 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 _ _ root _ _
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 _ _
```

Universal Dependencies

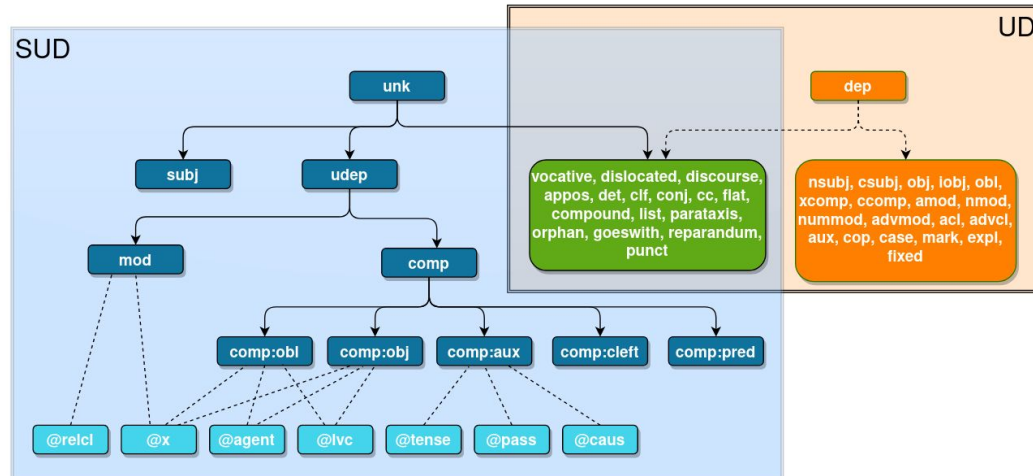
- Plusieurs projets d'annotation dans plusieurs langues
- Quelques projets multilingues
- En 2014, démarrage du projet UD
 - 10 corpus, 10 langues dans la version 1.0
 - 245 corpus, 141 langues dans 2.12
 - <https://universaldependencies.org/>





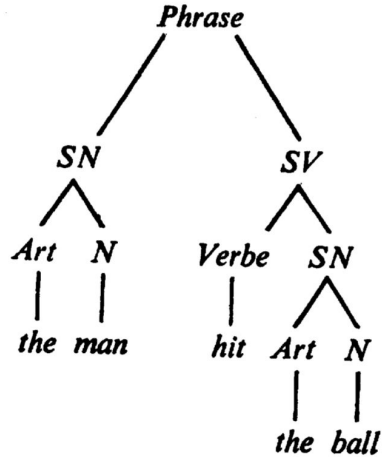
Surface Syntactic UD

- Alternative à l'UD
- Les relations sont définies sur des bases distributionnelles et fonctionnelles.
- <https://surfacesyntacticud.org/>

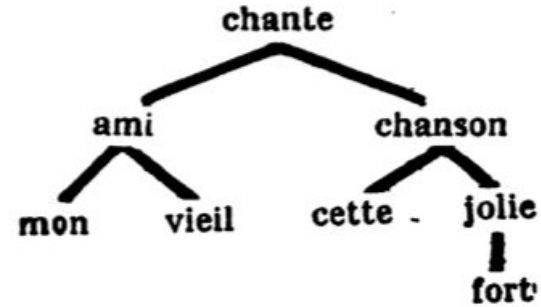


Structures et analyses

Analyse en constituants vs en dépendance



Arbre de constituants
Chomsky 1957 (version 1969)

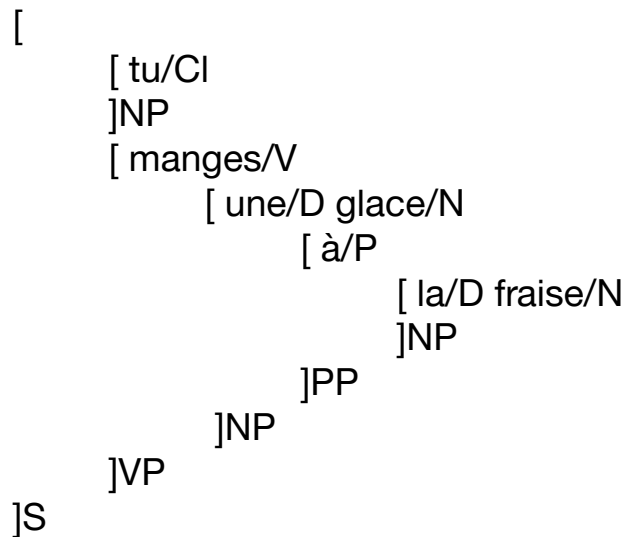
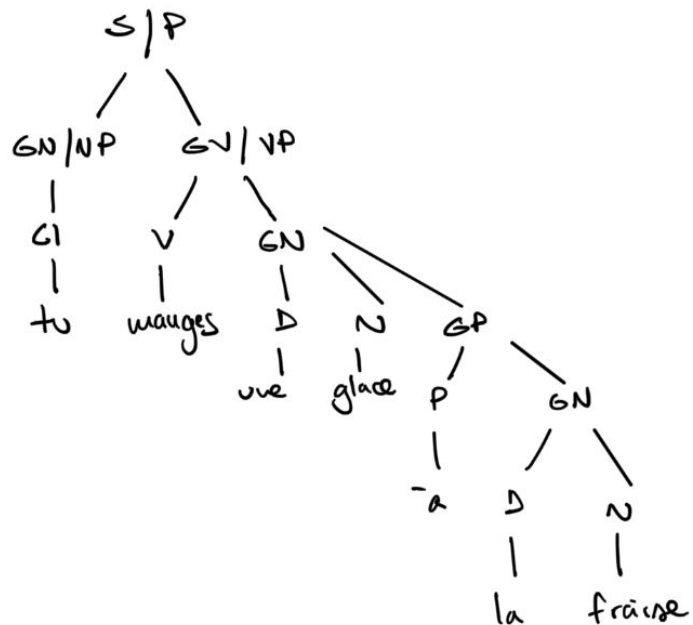


Arbre de dépendance
Tesnière 1959

Analyse en constituants

- Arbre de constituants = parenthésage

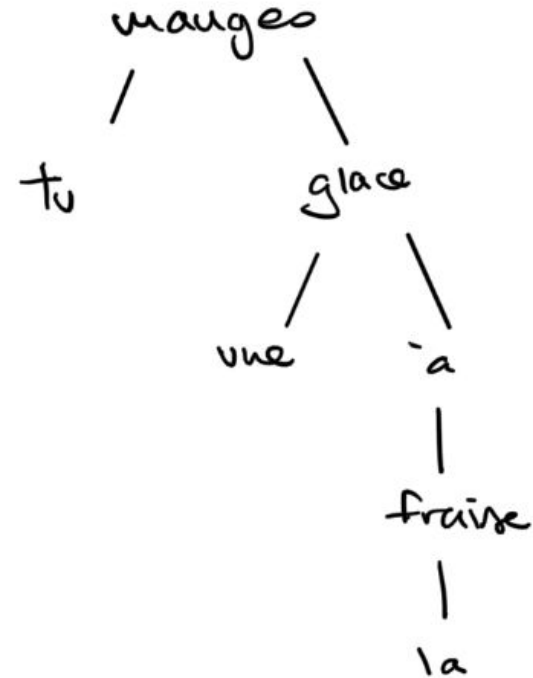
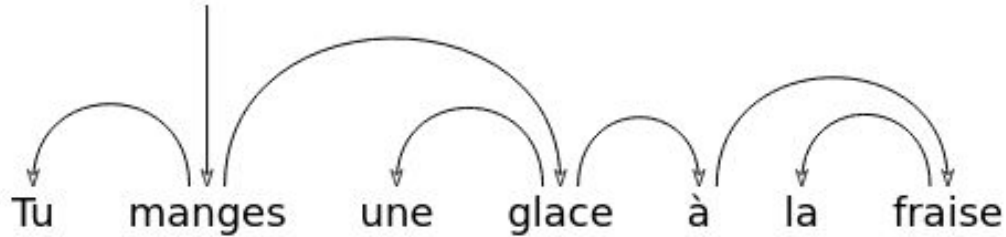
 Tu manges une glace à la fraise



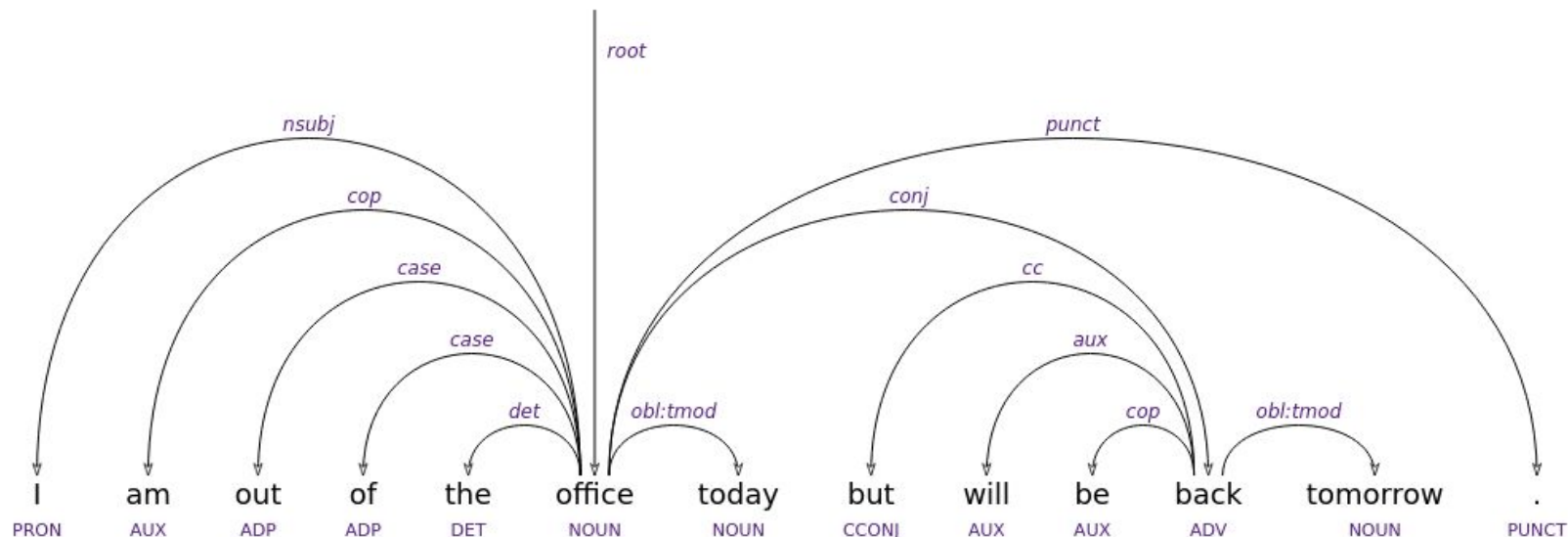
Analyse en dépendance

- Arbre en dépendance

 Tu manges une glace à la fraise

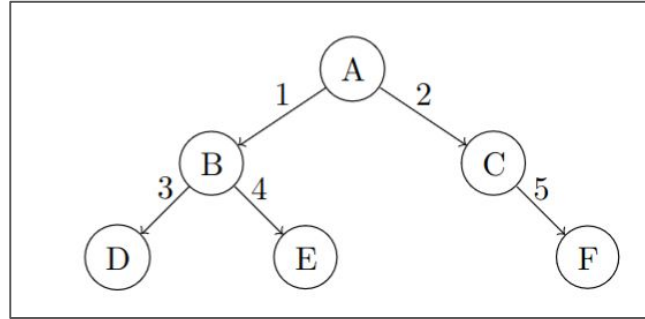


Analyse en dépendance

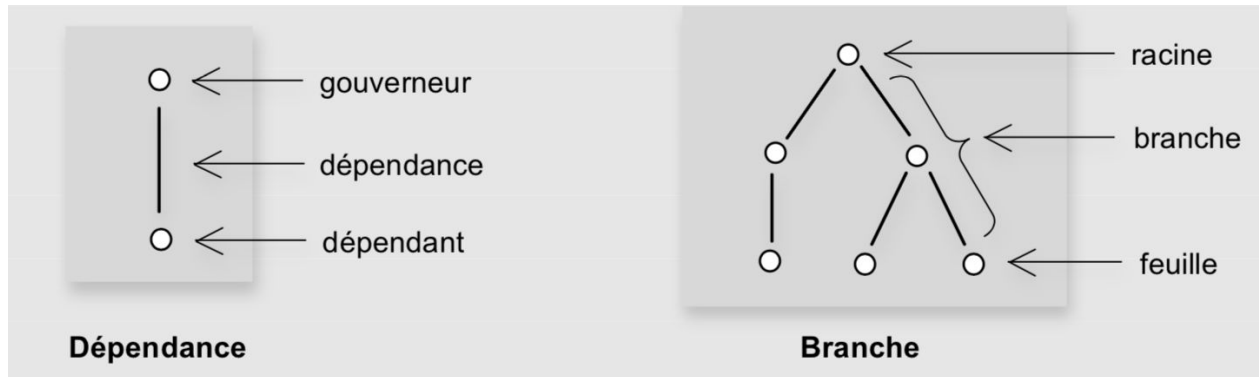


- Chaque **mot** est un **noeud** dans l'arbre
- Chaque **relation syntaxique** est une **arête** dans l'arbre
- Structure minimale : $n-1$ connexions pour n noeuds/mots
- Des outils disponibles : système de requêtes, algorithmes de parsing, etc.
- Évaluation des parsers plus simple

Un arbre enraciné et étiqueté



Quelques notions importantes :



Pourquoi fait-on des treebanks?

À l'ère pré-numérique :

- À des fins pédagogiques (trouver des exemples de constructions).
- À des fins théoriques (tester une théorie linguistique à l'aide d'exemples réels).

À l'époque pré-LLM :

- Comme entrée et sortie dans les outils de TAL : création et évaluation des parseurs, extraction d'information, traduction automatique
- Pour la recherche linguistique, l'extraction de grammaires

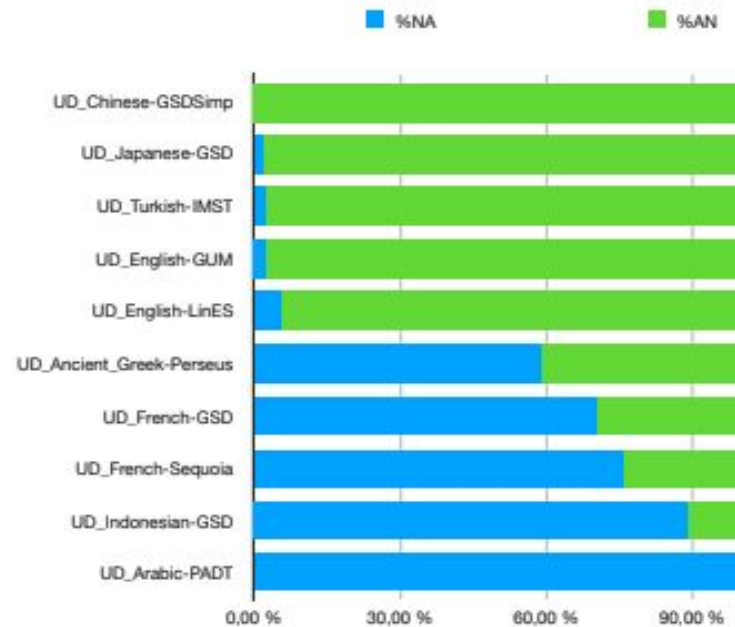
Après les LLMs :

- L'enseignement
- Évaluation des systèmes de TAL et de LLMs
- Dans des scénarios avec peu de données
- Pour la recherche linguistique (syntaxe, typologie, extraction de grammaires à partir de corpus)

A quoi servent les treebanks ?

Y a-t-il un ordre différent NOUN/ADJ en fonction des langues ? Si oui, où le français se situe-t-il ? Quelles sont les langues qui se comportent différemment ?

Corpus	NA	AN	%NA	%AN
UD_Chinese-GSDSimp	3	2028	0,15 %	99,85 %
UD_Japanese-GSD	8	462	1,70 %	98,30 %
UD_Turkish-IMST	70	3297	2,08 %	97,92 %
UD_English-GUM	157	6451	2,38 %	97,62 %
UD_English-LinES	235	4258	5,23 %	94,77 %
UD_Ancient_Greek-Perseu	710	496	58,87 %	41,13 %
UD_French-GSD	14139	6003	70,20 %	29,80 %
UD_French-Sequoia	2887	925	75,73 %	24,27 %
UD_Indonesian-GSD	4061	505	88,94 %	11,06 %
UD_Arabic-PADT	24335	86	99,65 %	0,35 %



Syntaxe 101 : Rappel des notions de base

Syntaxe

étude de la manière de laquelle les mots (et/ou les morphèmes) s'organisent pour former des unités plus larges (groupes/syntagmes, propositions, phrases)

Thématiques centrales :

- relations grammaticales
- ordre des mots
- structure hiérarchique de la phrase
- accord
- variation entre les langues

Syntaxe

Théories syntaxiques majeures :

- Grammaires en dépendances (L. Tesnière, Théorie Sens \leftrightarrow Texte)
- Grammaires catégorielles (Tree Adjoining Grammar de A. Joshi)
- Syntaxe générative (N. Chomski)
- Grammaires cognitives (Construction Grammar, e.g. Fillmore)

La représentation de la phrase et de sa structure varient d'une théorie à l'autre

Syntaxe : à l'intérieur d'une proposition simple

1. **Jean** mange une pomme

sujet: pronom. en **il/elle** ;
qui est-ce **qui** / qu'est-ce **qui**

2. Jean **mange** une pomme

verbe principal / racine
porteur du prédicat => détermine
les arguments

3. Jean mange **une pomme**

COD : pronom. en **le/la/les**
qui est-ce **que** / qu'est-ce **que**

Syntaxe : à l'intérieur d'une proposition simple

6. Jean mange **une** pomme

déterminant ; actualise le référent du nom

7. Jean mange une **belle** pomme

épithète ; exprime une propriété du nom

8. Jean donne une pomme **à Marie**

COI; bénéficiaire ou expérienceur
pronom. en **lui**; à **qui/quoi**

Syntaxe : à l'intérieur d'une proposition simple

7. Jean parle **avec Marie**

complément circonstanciel
d'accompagnement

8. Jean conduit **avec prudence**

complément circonstanciel de
manière

9. Jean parle avec Marie **à l'école**

complément circonstanciel de lieu

10. Jean va **à l'école**

complément obligatoire de lieu

non obligatoire => **modifieur**

obligatoire => **complément**

Syntaxe : à l'intérieur d'une proposition simple

- | | |
|--|--|
| 11. La belle pomme de Jean | complément de nom ; typiquement un groupe prépositionnel |
| 12. La belle pomme de Jean | complément de préposition ; typiquement un groupe nominal |
| 13. Jean rêve d'une belle pomme | COI ; pronom. en en |
| 14. Jean pense à la pomme | COI ; pronom. en y |

Syntaxe : subordination

1. Jean mange une pomme **qui est rouge** **relative** ; typiquement exprime la propriété d'un nom ; ≈ épithète
2. Jean dit **qu'il aime les pommes** **complétive** ; typiquement exprime l'objet direct d'un verbe de la parole ; ≈ COD
3. Jean mange des pommes **quand il veut** **circonstancielle / adverbiale** ; typiquement exprime les circonstances de la réalisation du prédicat de la principale ; ≈ CC

Syntaxe : subordination

4. Jean mange des pommes **parce qu'il les aime**

circonstancielle de cause
=> **causale**

5. Jean mange des pommes **pour être en bonne santé**

circonstancielle de but
=> **finale**

6. Jean aime tellement les pommes **qu'il en mange tout le temps**

circonstancielle de conséquence
=> **consécutive**

7. Jean mange des pommes **bien qu'il en ait marre**

circonstancielle de concession
=> **concessive**

Syntaxe : coordination

1. **Jean et Marie** parlent coordination de **sujet**
2. Jean **parle et rit** coordination de **verbe principal**
3. Jean mange des pommes **rouges et vertes** coordination d'**épithète**
4. Jean mange **des pommes rouges et des pommes vertes** coordination de **COD**

Syntaxe : coordination

- 5. **Jean mais pas Marie** est venu coordination de **sujet**

- 6. Jean est venu **mais Marie a renoncé** coordination de **propositions**

Syntaxe : coordination

- | | |
|---|---|
| 7. Jean est étudiant et très content | conjointes de nature différente |
| 8. Jean fait des crêpes et Marie des madeleines | ellipse : gapping |
| 9. Jean fait et Marie mange des crêpes | ellipse : conjoints qui ne sont pas des constituants |

Allez sur le site UD et choisissez un treebank d'une langue que vous ne connaissez pas. Trouvez des informations sur le treebank et soyez critiques sur la taille, la source des phrases, les choix d'annotation etc. À quoi pourrait servir ce treebank?

À l'ère des LLMs, quel rôle unique les treebanks peuvent-ils encore jouer, que les grands modèles ne peuvent totalement remplacer?

Si vous étiez en train de créer un nouveau treebank pour une langue sous dotée, quel cadre théorique et quel schéma d'annotation choisiriez-vous (dépendants vs. constituants, UD vs SUD, autre chose?), et pourquoi?

À la maison

En utilisant les notions de syntaxe présentées, annotez les phrases suivantes:

1. Pierre mange une pomme et boit un café avec Marie.
2. Marie lit des romans et des poèmes dans le jardin.
3. Pierre donne une pomme à Jean qui est fatigué.
4. Paul aime les crêpes parce qu'elles lui rappellent son enfance.
5. Pierre veut acheter et lire un nouveau livre pour ses recherches.

Challenge: Faites les arbres de constituants / de dépendance pour ces phrases.

À la maison - exemple

Marie	mange	une pomme	à l'école.
sujet	verbe	COD	CC
	principal		de lieu

