Sorbonne Nouvelle
université des cultures

inalco
Institut national des langues et civilisations orientales

Université Paris Nanterre

Corpus arborés et parsing

M1 - pluriTAL

Aleksandra Miletić - Chercheuse
Santiago Herrera - Doctorant en deuxième année

Mon parcours

Modèles, Dynamiques, Corpu
UMR 7114

- Licence en langue et littérature françaises (Belgrade, Serbie)
- Master LTTAC (Lille) : création de ressources pour la linguistique
- Doctorat (Toulouse) : Un treebank pour le serbe : constitution et exploitation
- Depuis : constitution d'outils et de ressources pour les langues et variétés linguistiques moins dotées

Votre tour

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

Cours

- 12 séances
  - 6 séances avec Aleksandra
  - 6 séances avec Santiago

- Besoin d'un ordinateur
- Concepts théoriques et exercices pratiques
- Plateforme CoursEnLigne

Évaluation

- DM notés
  - 2023/2024 : 3 DMs répartis sur le semestre
- Exercices en ligne
  - Réalisation collaborative d’un treebank en ligne
- Examen présentiel ou DM final
  - A définir

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’ vient du latin : corps et collection
- Un ensemble de textes ou discours produits/attestés

- 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 .

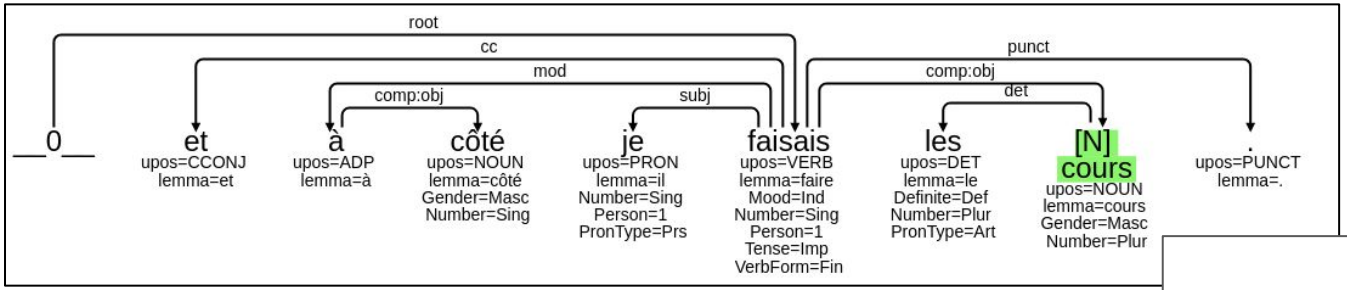
- Un bon corpus doit comporter des métadonnées
- Jeu de données (*dataset*) vs Corpus
- Format numérique

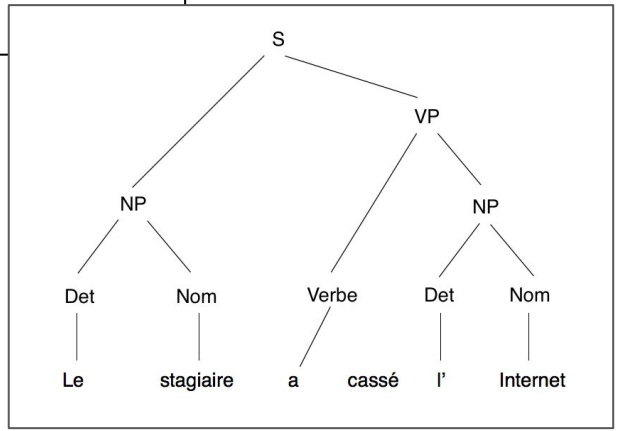
De quoi parlons-nous lorsque nous parlons de …

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 **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 du langage naturel (ou d’une autre unité de segmentation) afin de trouver sa structure et de catégoriser ses éléments.
  - ✏️ Parsez **I saw a woman with the telescope wrapped in paper**

- En **psycholinguistique :** Le parsing est souvent considéré incrémental et implique l'analyse et la compréhension d’un énoncé.

De quoi parlons-nous lorsque nous parlons de …

Création des treebanks ou des parsebanks

- Un parseur s'entraîne avec des treebanks
- On peut commencer par des annotations automatiques

Plan succinct

- 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)
- Théorie sur le parsing et bootstrapping

Brève histoire des diagrammes syntaxiques

Diagrammes syntaxiques

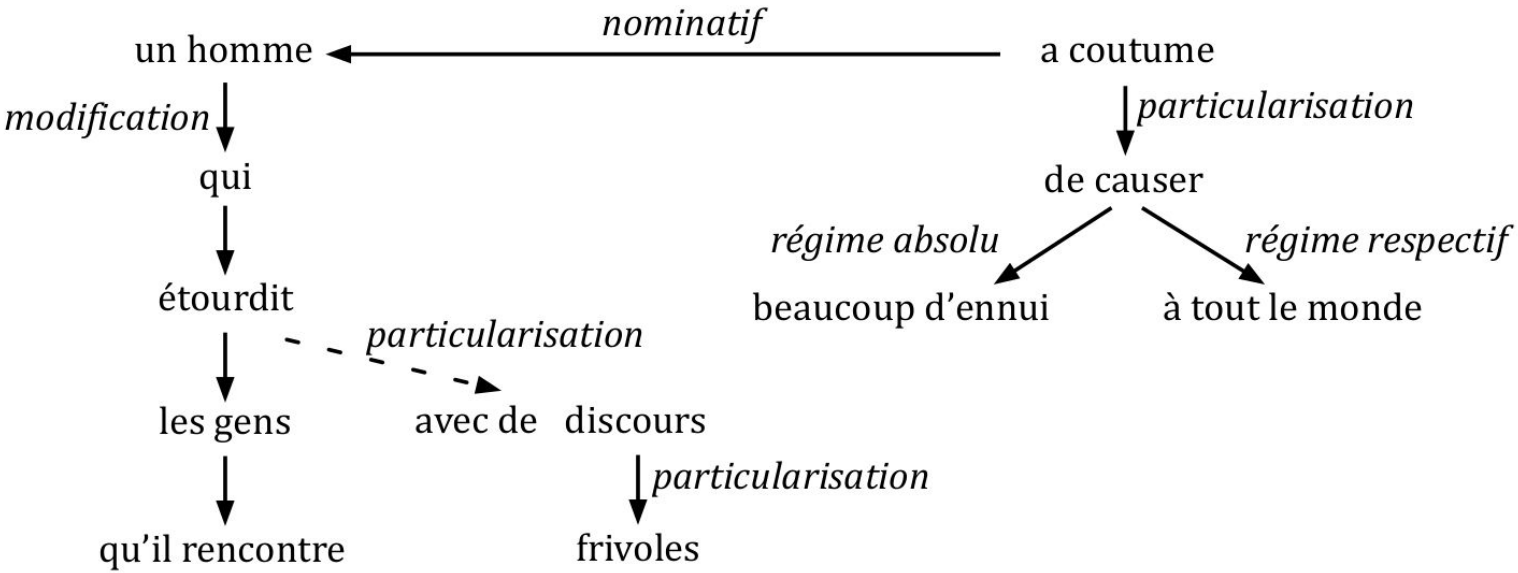
Claude Buffier (1709)

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 l**es 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.**

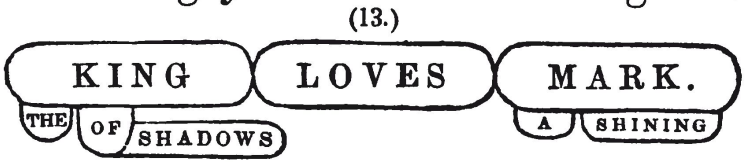
Voilà donc comment tous les mots d'une phrase quelque longue qu'elle soit, ne sont que pour modifier le nom & le verbe.

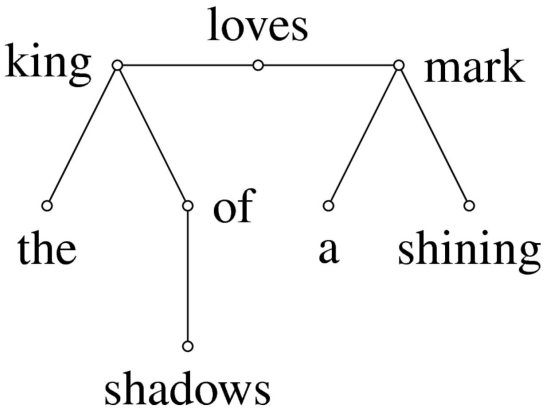
Diagrammes syntaxiques : Claude Buffier (1709)



Diagrammes syntaxiques : Stephen W. Clark (1847)

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

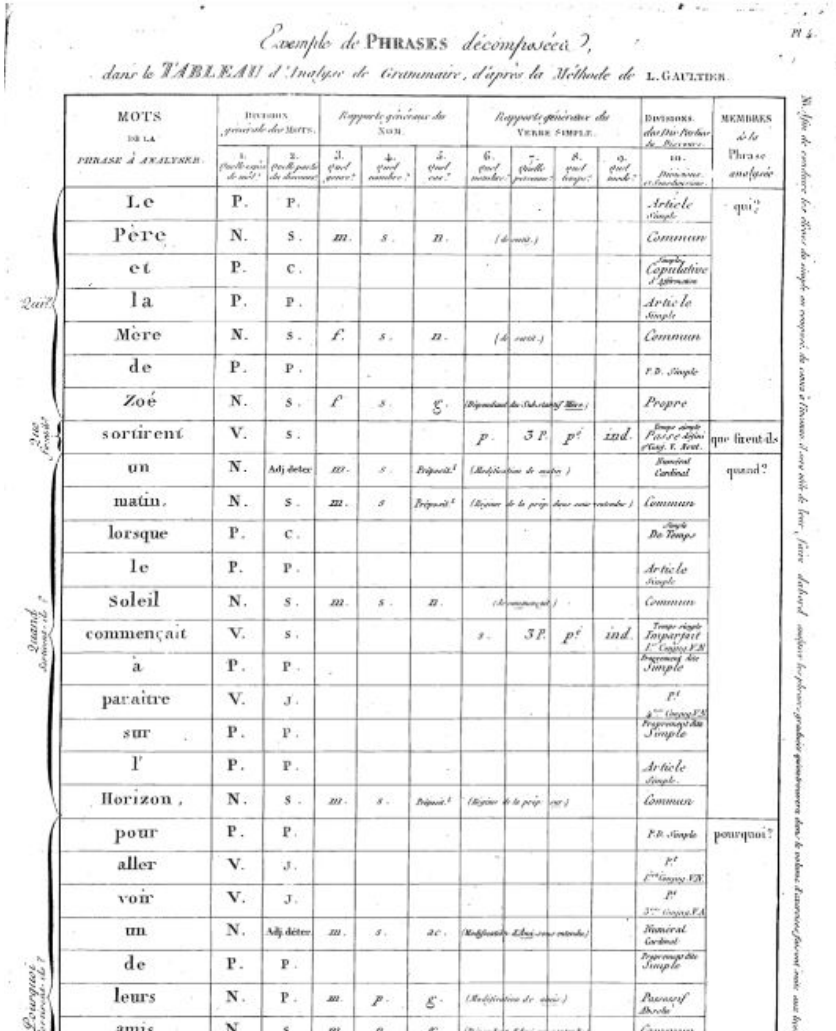




Une structure de dépendance dé-réifiée

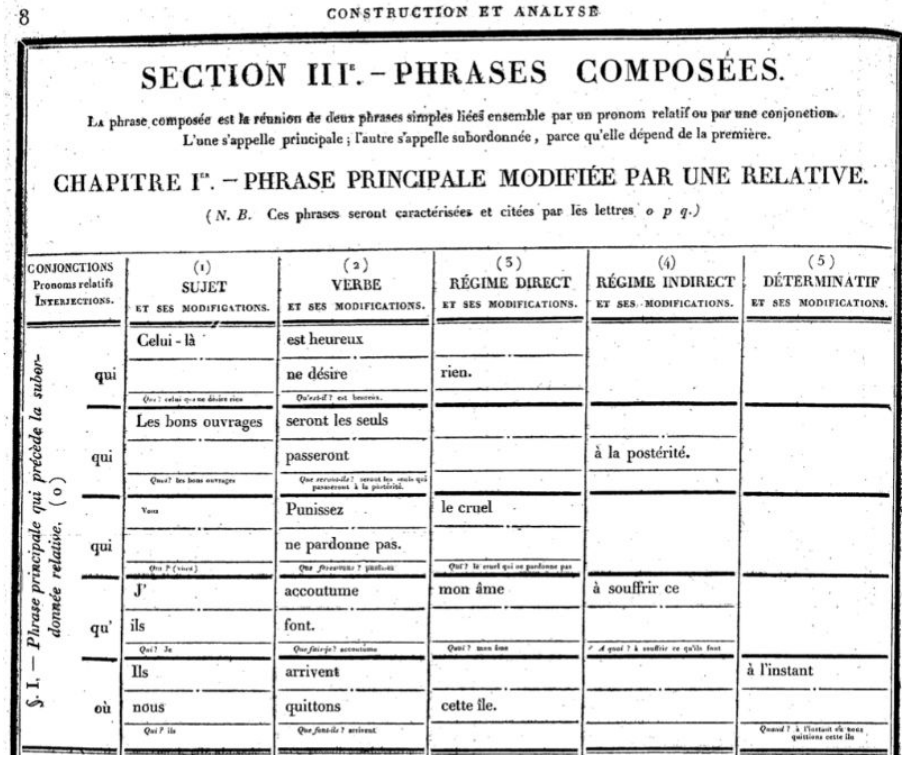
Diagrammes tabulaires

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.



Diagrammes tabulaires

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



8 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.

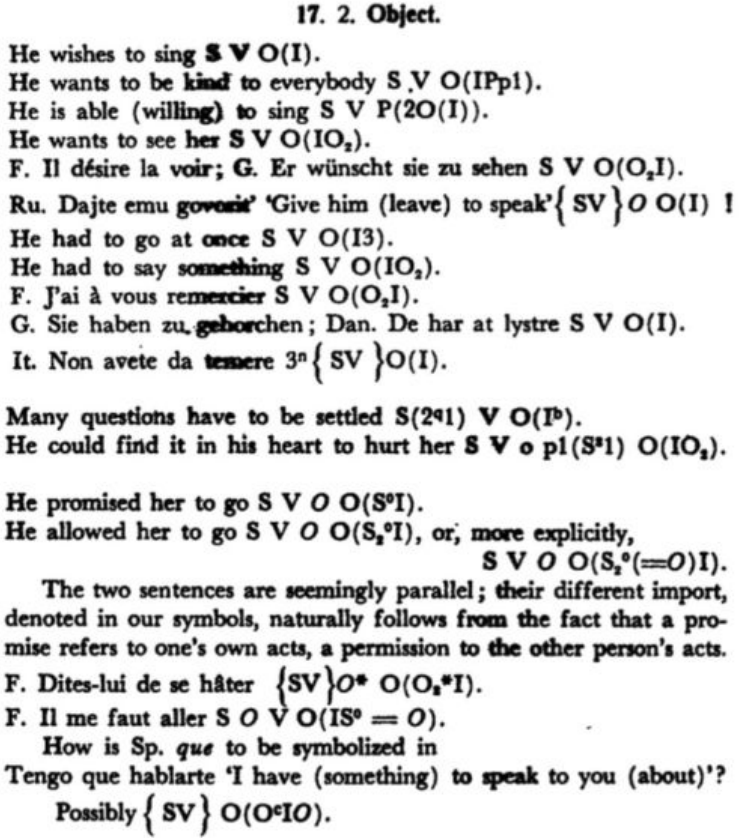
CHAPITRE I^er. – PHRASE PRINCIPALE MODIFIÉE PAR UNE RELATIVE.

(*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.		
	Qui? celui-là ne désire rien	Qu'est-il? est heureux.			
qui	Les bons ouvrages	seront les seuls			
		passeront		à la postérité.	
	Quoi? les bons ouvrages	Que reviendra? seront les seuls qui passeront à la postérité.			
qui	Vous	Punissez	le cruel		
		ne pardonne pas.			
	Qui? (vous)	Que ferez-vous? punissez	Qui? le cruel qui ne pardonne pas		
qu'	J'	accoutume	mon âme	à souffrir ce	
	ils	font.			
	Qui? Je	Que faites-je? accoutume	Qui? mon âme	A qui? à souffrir ce qu'ils font	
où	Ils	arrivent			à l'instant
	nous	quittons	cette île.		
	Qui? ils	Que font-ils? arrivent			Quand? à l'instant où nous quittons cette île

Diagrammes : analyse en constituants

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



17. 2. Object.

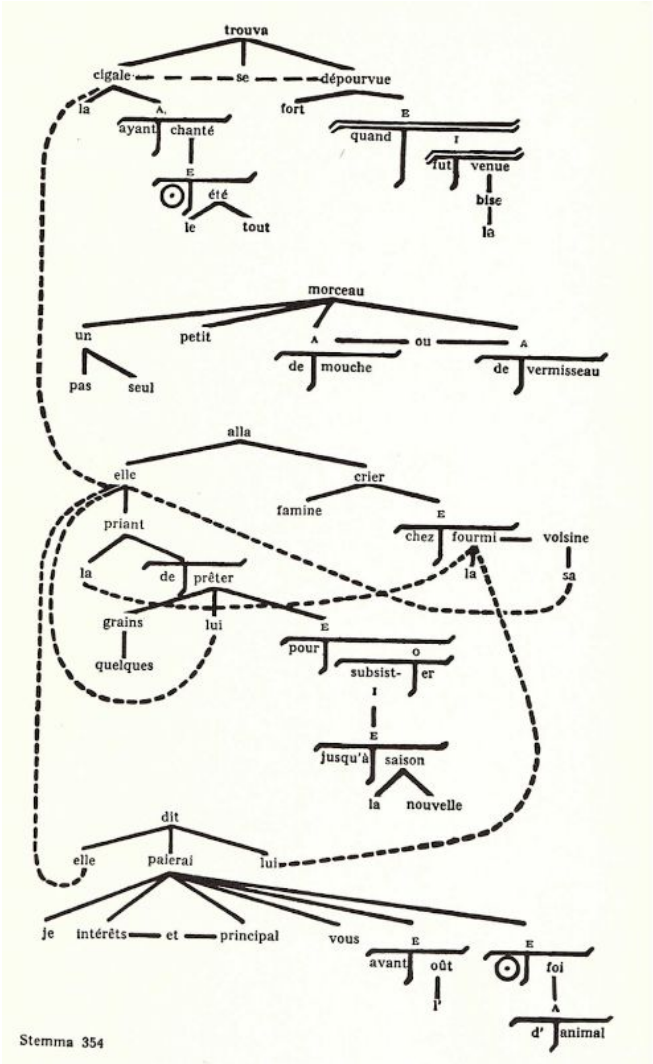
He wishes to sing **S V** O(I).
He wants to be **kind** to everybody S V O(IPp1).
He is able (**willing**) to sing S V P(2O(I)).
He wants to see **her** S V O(IO_2).
F. Il désire la **voir**; G. Er wünscht sie zu sehen S V O(O_2I).
Ru. Dajte emu **govorit'** 'Give him (leave) to speak' { SV } *O* O(I) !
He had to go at **once** S V O(I3).
He had to say **something** S V O(IO_2).
F. J'ai à vous **remercier** S V O(O_2I).
G. Sie haben zu **gehorchen**; Dan. De har at lystre S V O(I).
It. Non avete da **temere** 3^n { SV } O(I).

Many questions have to be settled S(2^{q}1) **V** O(I^{b}).
He could find it in his heart to hurt her **S V** o p1(S^{s}1) O(IO_3).

He promised her to go S V *O* O(S^{o}I).
He allowed her to go S V *O* O(S_2^{o}I), or, more explicitly,
S V *O* O(S_2^{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(O_2*I).
F. Il me faut aller S *O* V O(IS^{o} = *O*).
How is Sp. *que* to be symbolized in
Tengo que hablarte 'I have (something) to speak to you (about)'?
Possibly { SV } O($O^{c}IO$).

Diagrammes en dépendance

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

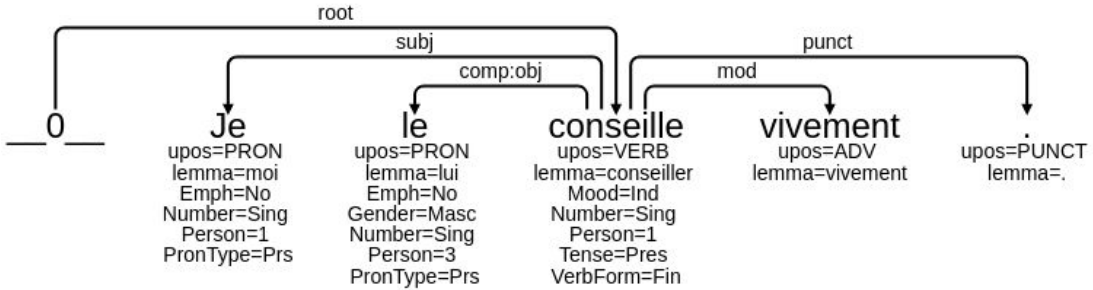


S'agit-il de treebanks ?

- 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: Dependency parsing becomes dominant
- 2005: the Stanford parser (2002) proposes a dependency-based output
- 2007: CoNLL dataset => **CoNLL** format for dependency trees
- 2008: POS interset, many projects of conversion
- 2014: Google provides treebanks for 30 languages (based on Stanford schema)
- 2014: **Universal Dependencies** starts

Treebanks aujourd'hui

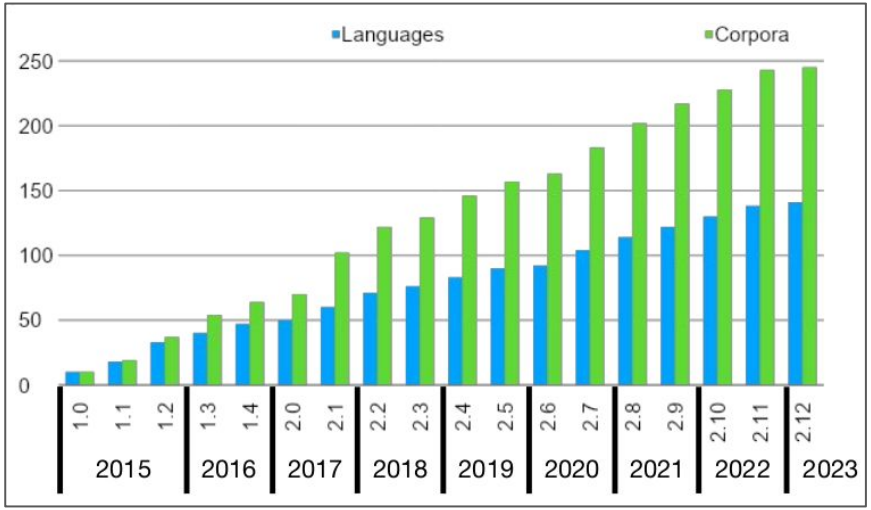
- Format numérique, requêtable et encodé dans un conllu
- Annotation simplifiée




```
# 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   _   _
```

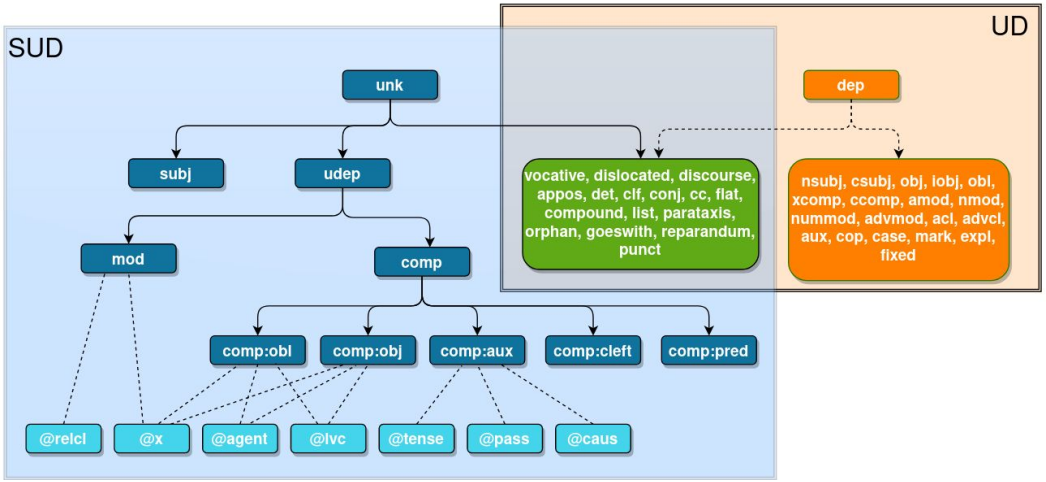
Universal Dependencies

- 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



Surface Syntactic UD

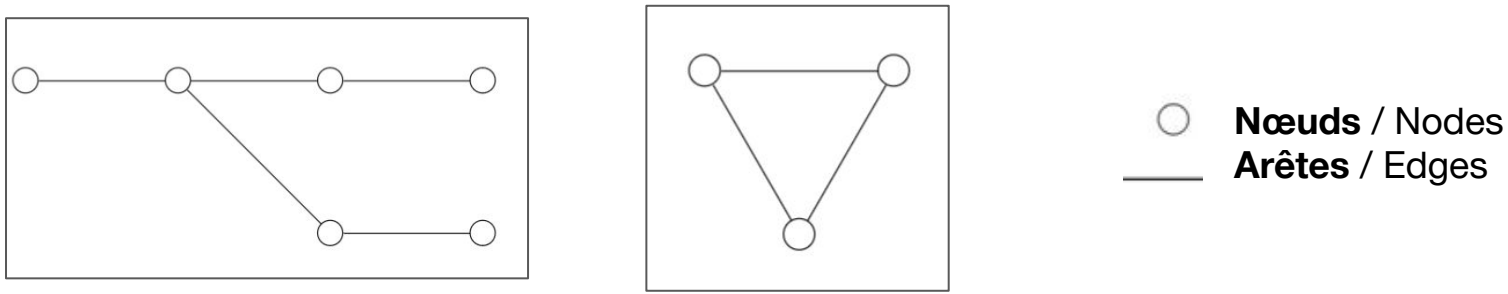
- Alternative à l'UD
- Basés sur des critères distributionnels
- Les relations sont définies sur des bases distributionnelles et fonctionnelles.



Structures et analysis

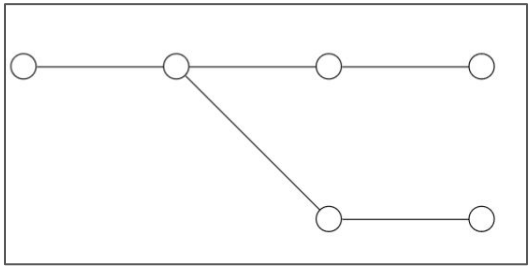
Pourquoi des arbres ?

- Une structure mathématique et de données
  - hiérarchique
  - capable de modéliser de relations entre éléments
- Plus précisément, un type de graphe

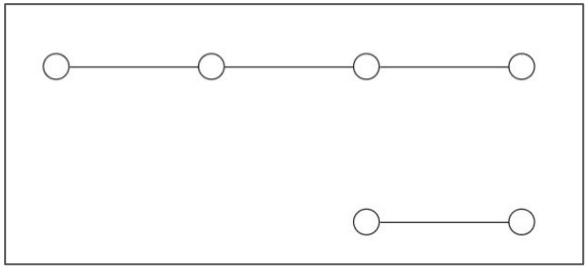


Graphes

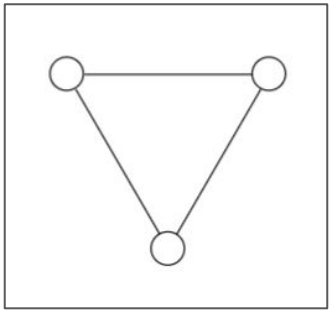
Graphes



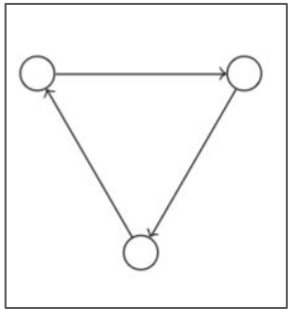
Graphe **connexe acylique**



Graphe **non connexe acylique**

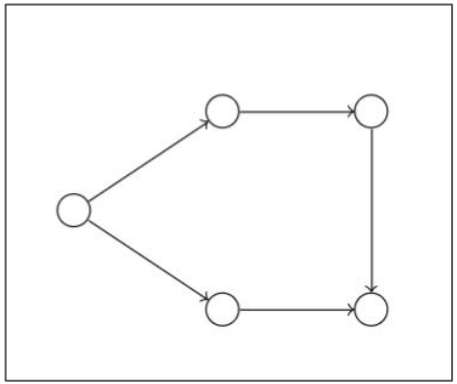


Graphe **cyclique**

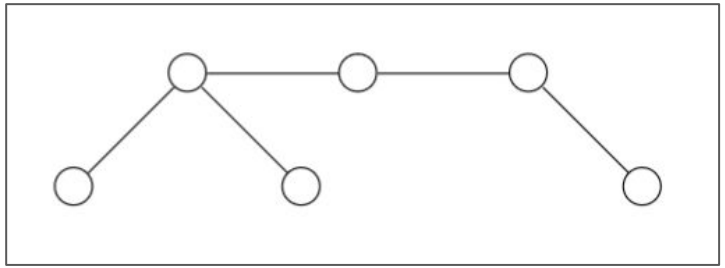


Graphe **cyclique orienté**

DAGs et Arbres

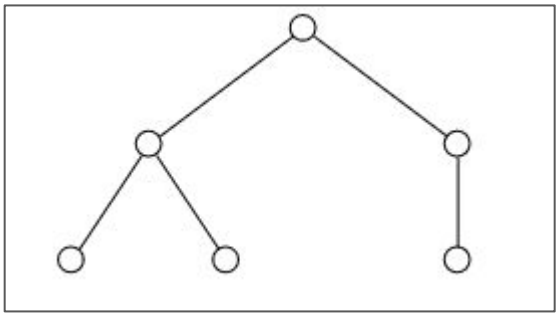


DAG / Graphe orienté acyclique



Arbre (un type de graphe connexe)

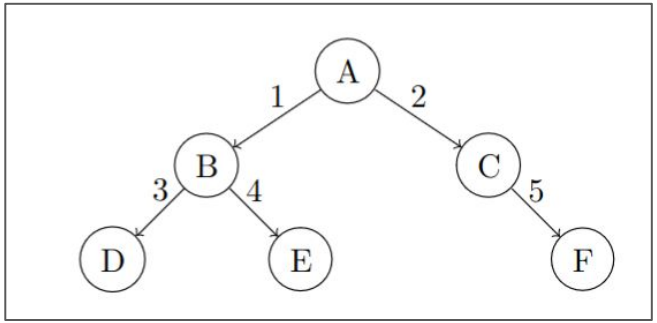
- faiblement connexe
- acyclique
- un seul chemin entre deux nœuds



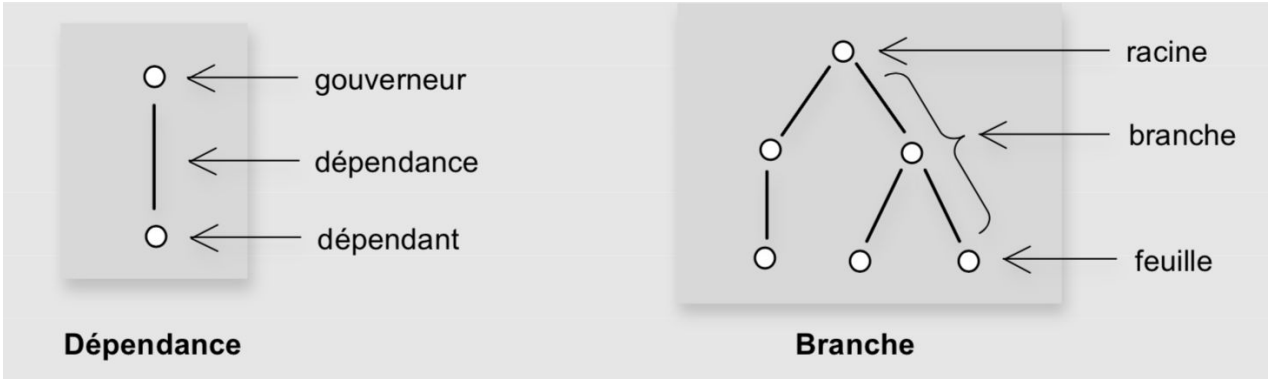
Arbre enraciné, alors orienté/pointé

- avec une racine
- chaque nœud est cible d'une et une seule arête (sauf la racine)

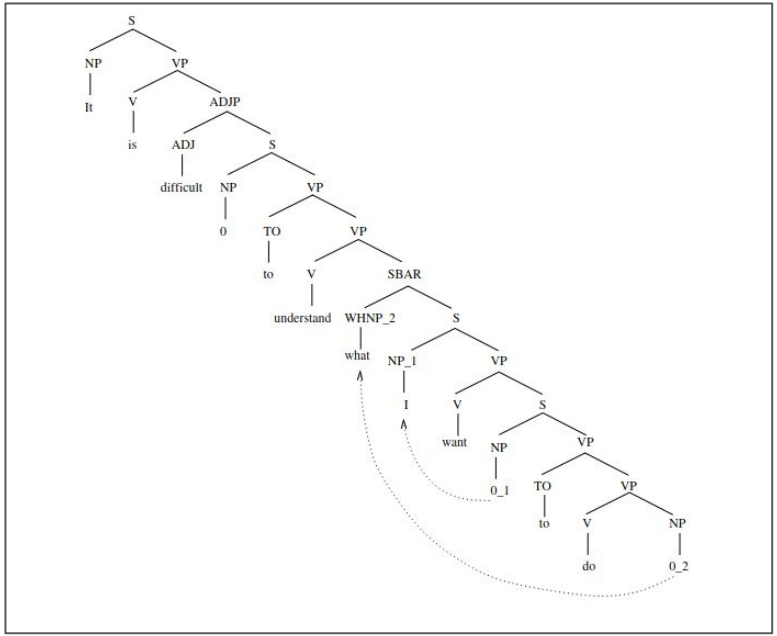
Un arbre enraciné et étiqueté



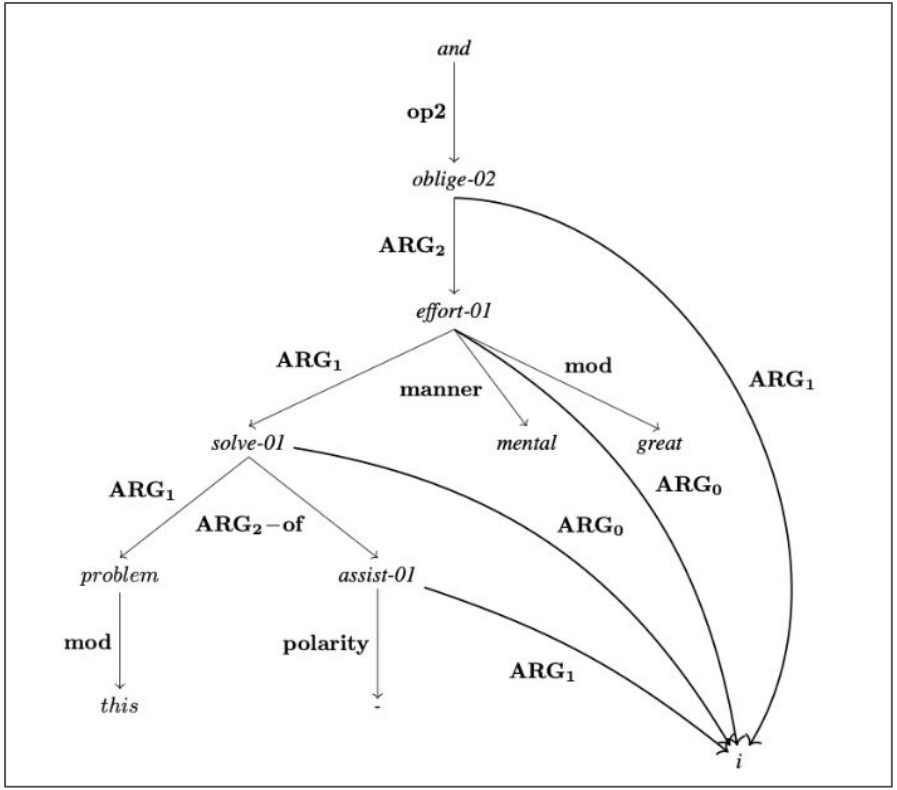
Quelques notions importantes :



DAGs

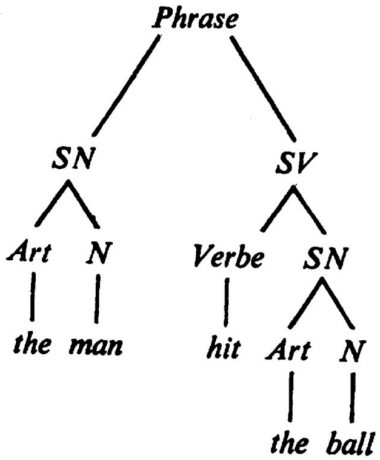


First example in Clark's "Penn Treebank Parsing"

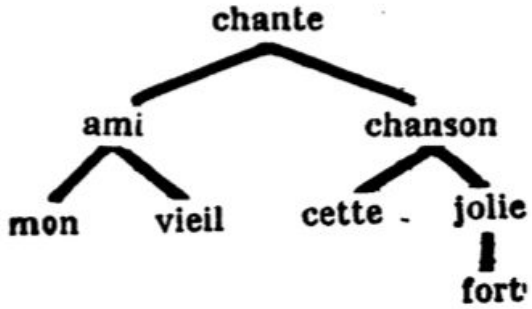


AMR pour *And I was obliged to make a great mental effort to solve this problem, without any assistance*

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

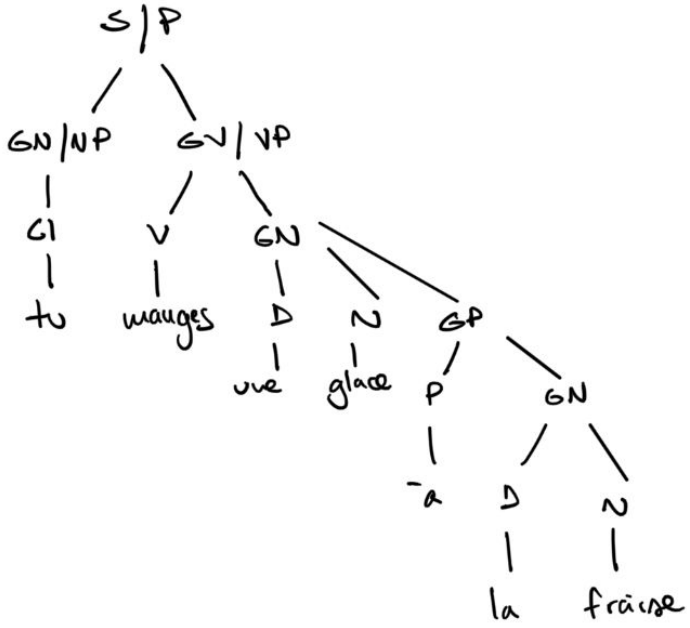
✏️ Tu manges une glace à la fraise

Analyse en constituants

- Arbre de constituants = parenthésage

✏️ Tu manges une glace à la fraise

```
[
        [ tu/Cl
        ]NP
        [ manges/V
                [ une/D glace/N
                        [ à/P
                                [ la/D fraise/N
                                ]NP
                        ]PP
                ]NP
        ]VP
]S
```



Analyse en dépendance

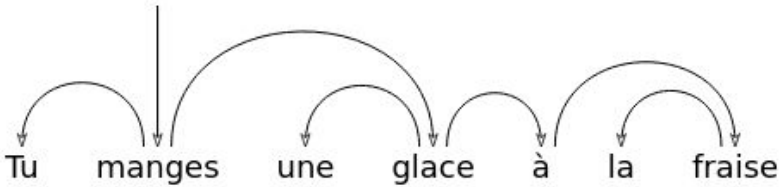
- Arbre en dépendance

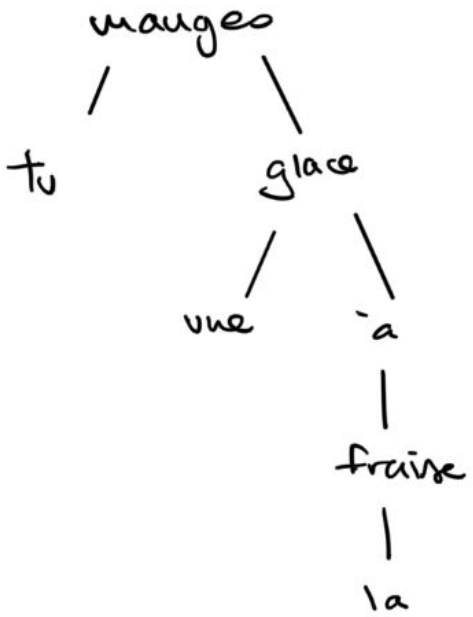
✏️ Tu manges une glace à la fraise

Analyse en dépendance

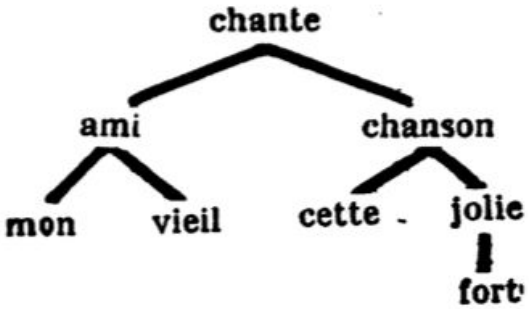
- Arbre en dépendance

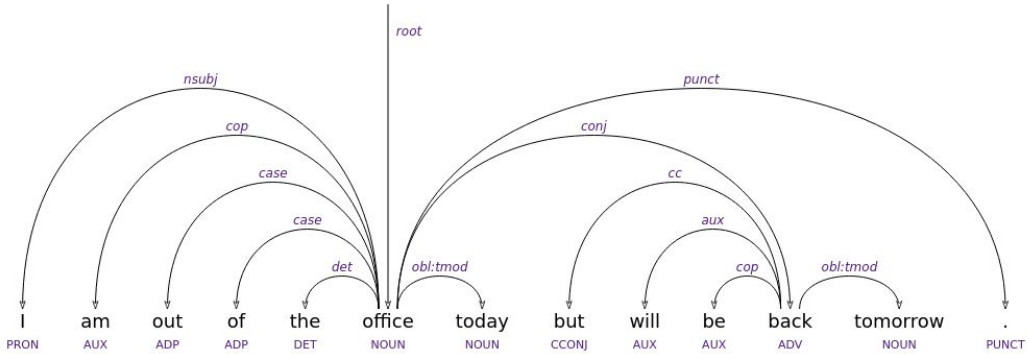
✏️ Tu manges une glace à la fraise





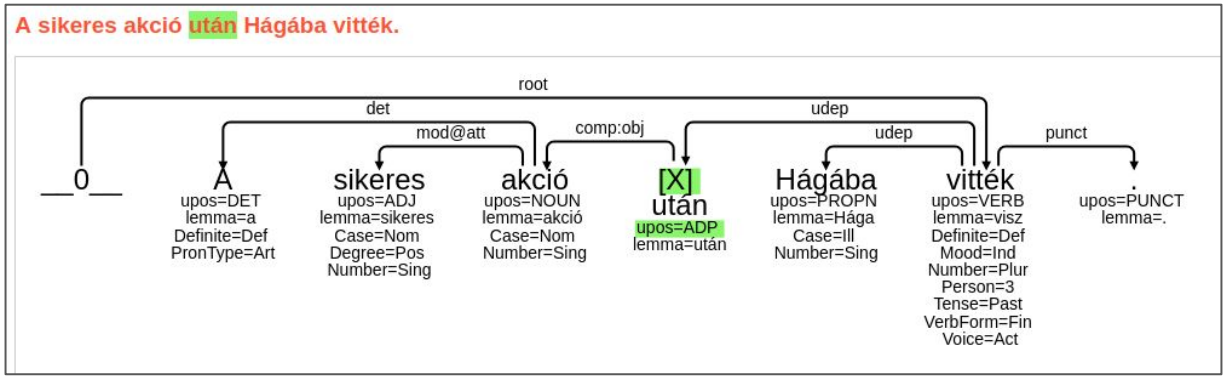
Analyse en dépendance

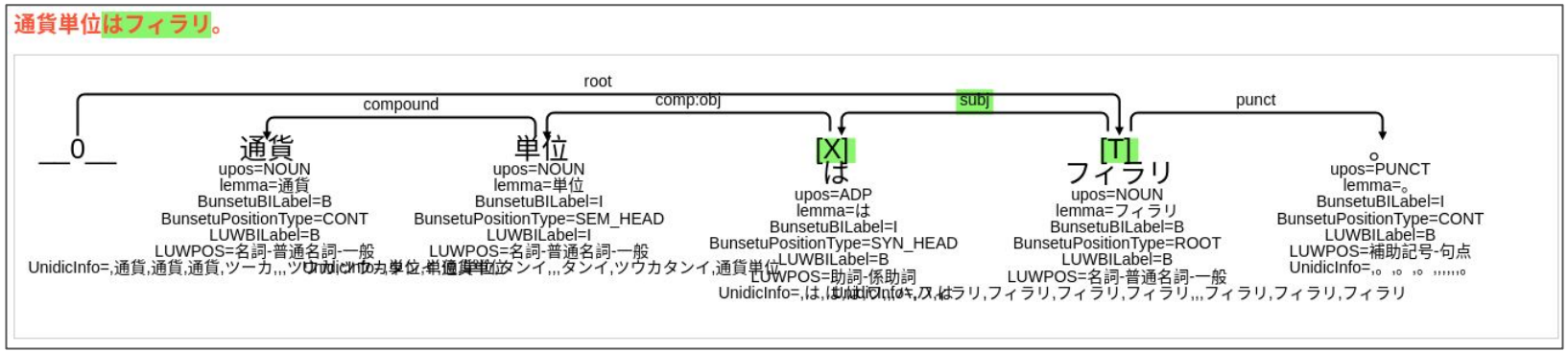




- Structure minimale : *n-1* connexions pour *n* nœuds/mots
- Annotation plus rapide
- Des outils disponibles : système de requêtes, algorithmes de parsing, etc.
- Évaluation des parsers plus simple

Analyse de la structure syntaxique





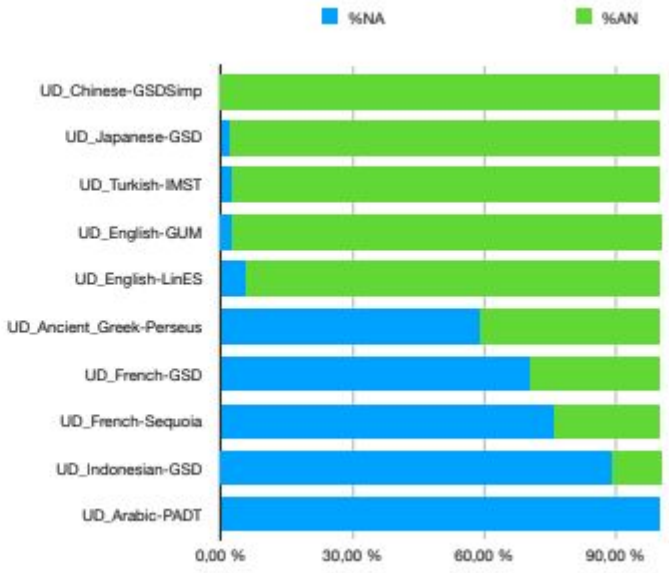
Analyse de la distribution d’une unité

Y'a t'il un ordre différent NOUN/ADJ en fonction des langues ? Si oui, ou se situe le français ? Quels sont les langues opposées ?

Analyse de la distribution des unités

Y'a t'il un ordre différent NOUN/ADJ en fonction des langues ? Si oui, ou se situe le français ? Quels sont les langues opposées ?

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 %



Pourquoi on fait 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 obtenir de la robustesse dans certaines tâches de TAL
- Pour la recherche linguistique
  (syntaxe, typologie, extraction de grammaires à partir de corpus)