Sorbonne Nouvelle
université des cultures

inalco
Institut national des langues et civilisations orientales

Université Paris Nanterre

Corpus arborés et parsing

Cours 2

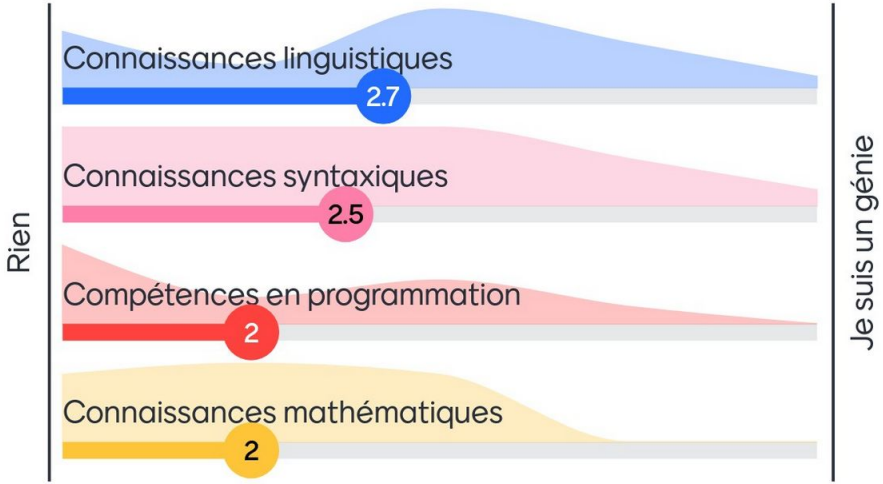
Santiago Herrera
s.herrera@parisnanterre.fr

Lien

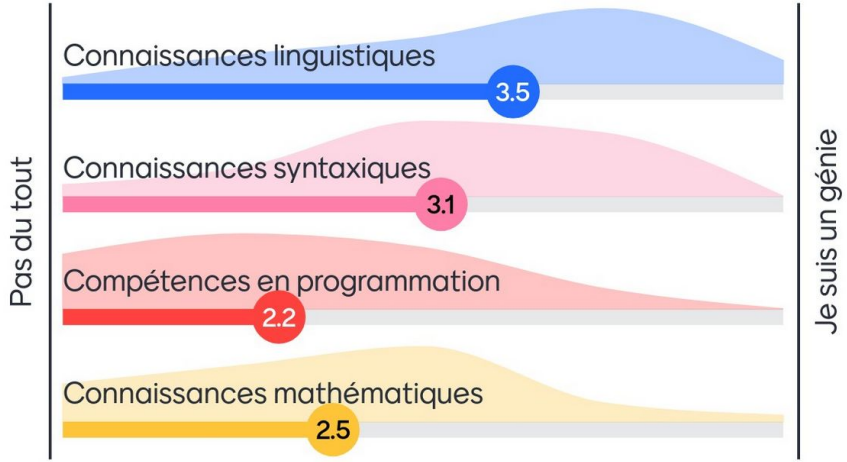
https://drive.google.com/drive/folders/1StwDH87QKnDppzVwkE7jrGJ2VbHNmB3f?usp=drive_link

Sondage : promo pluriTAL 2023

Auto-évaluation



Groupe 1



Groupe 2

Sondage : promo pluriTAL 2023

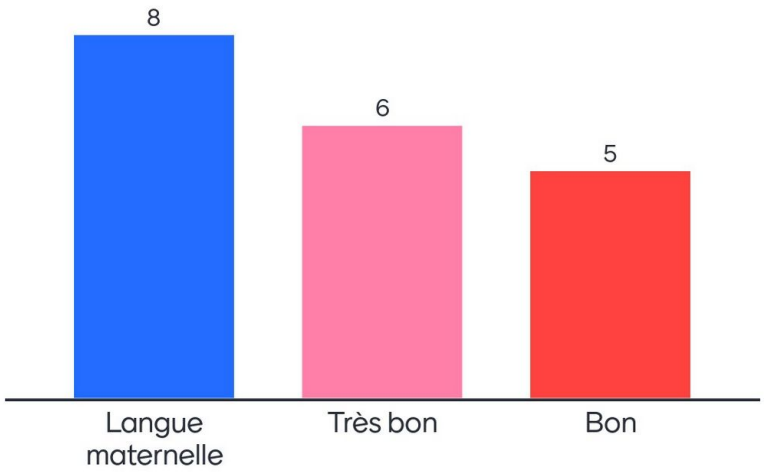
Langues parlées

Groupe 1

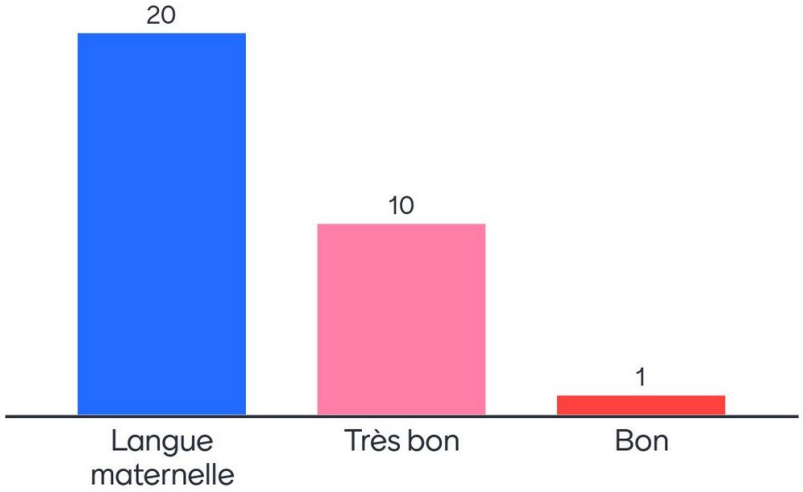
Groupe 2

Sondage : promo pluriTAL 2023

Niveau du français



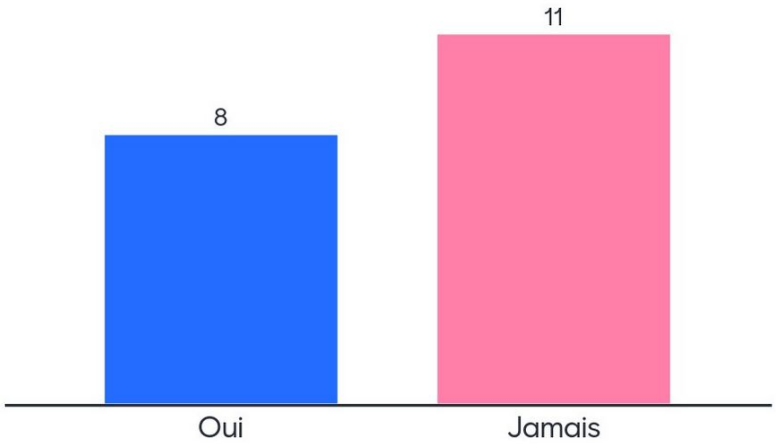
Groupe 1



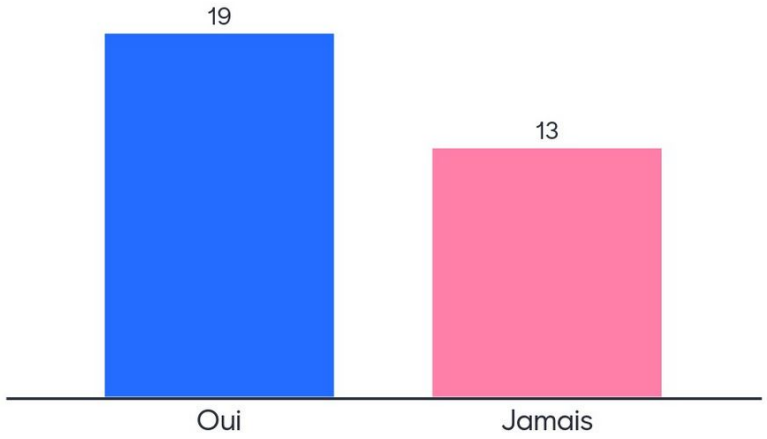
Groupe 2

Sondage : promo pluriTAL 2023

Avez-vous fait de l'analyse syntaxique en dépendance ?



Groupe 1

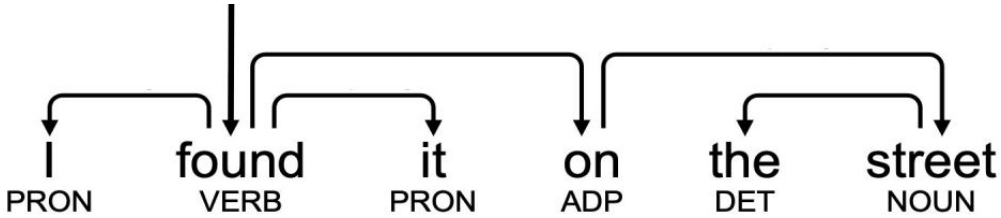


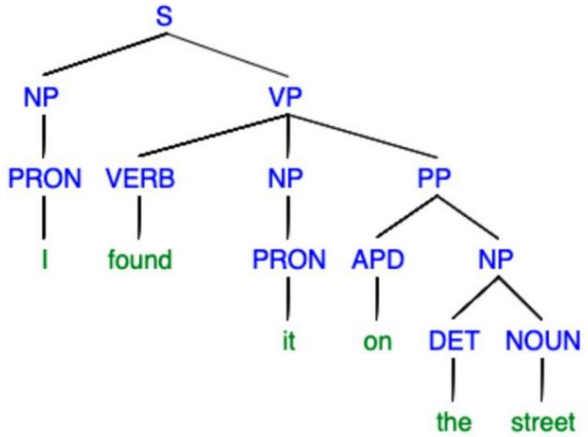
Groupe 2

Rappel : treebank

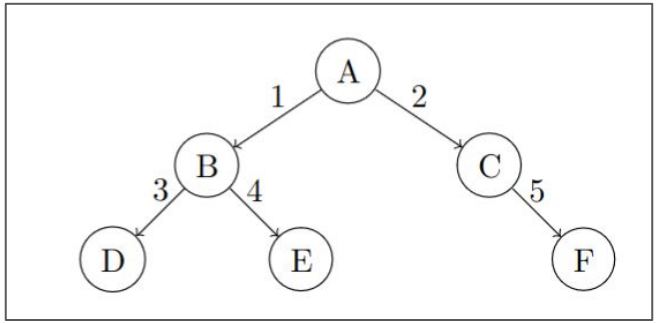
Un treebank est une collection de phrases associées à des arbres syntaxiques

- **Dans un treebank en dépendance**
  - Les connexions syntaxiques sont exprimées entre mots (unité d'analyse) vs. arbres de structure des phrases

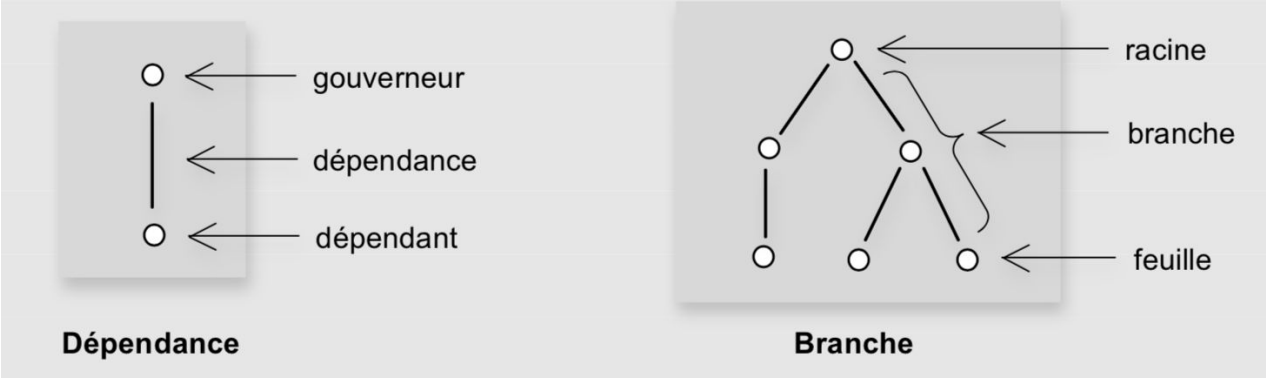




Rappel : un arbre enraciné et étiqueté



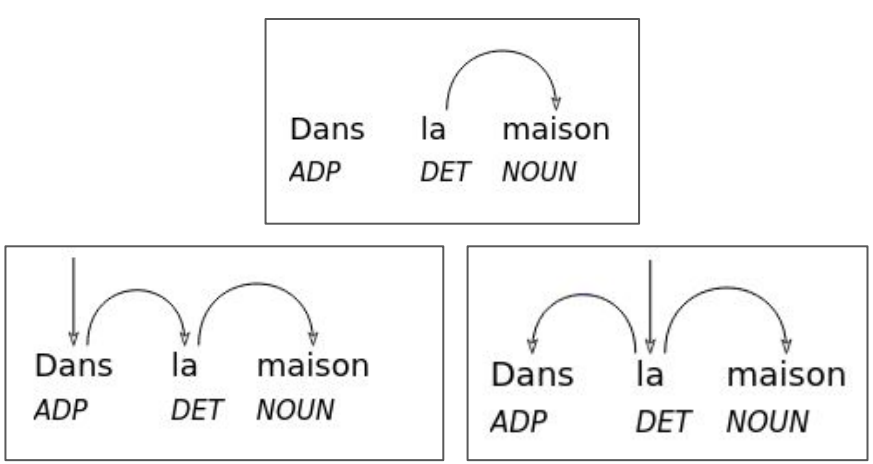
Et des notions importantes :

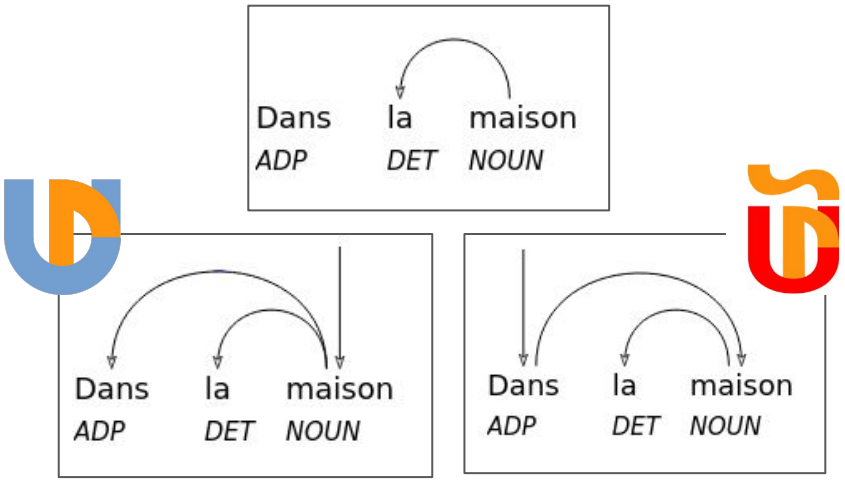


Rappel : contraints dans l'analyse en dépendance

1) Un mot va être **racine** de la phrase
2) Chaque **mot dépendent exactement d'un seul mot**

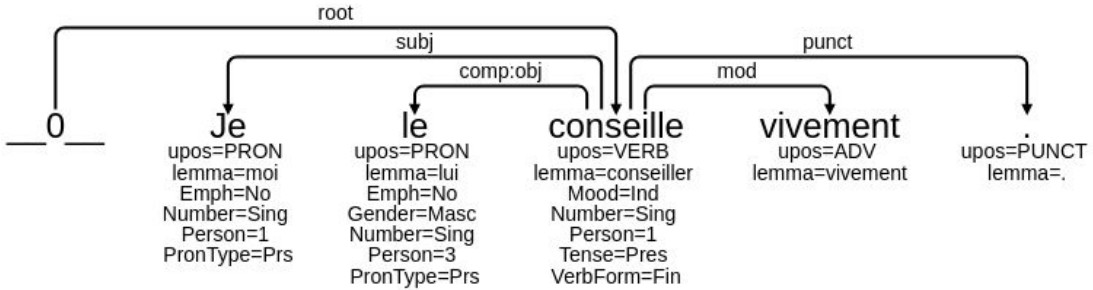
Malgré ces contraints, nous avons des choix d'analyse





Rappel : treebanks aujourd'hui

- Format numérique, requêtable et encodé dans un conll-u
- 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   _   _
```

Syntaxe

Définitions tirées du livre "Morphosyntax" (2022) de William Croft :

- **Syntax** is the analysis of the internal structure of utterances/**sentences** – more specifically, how **words** are put together.
- **Morphology** is the analysis of the internal structure of words, including prefixes, suffixes, and other internal changes to words that generally have a meaning.
- The term **morphosyntax** refers to the combination of morphology and syntax.

Mots

Il n'existe pas de définition canonique et universelle du mot :

- **Defining the word** (Haspelmath 2022): A word is (i) a free morph, or (ii) a clitic, or (iii) a root or compound plus possibly nonrequired affixes plus required affixes if there are any.

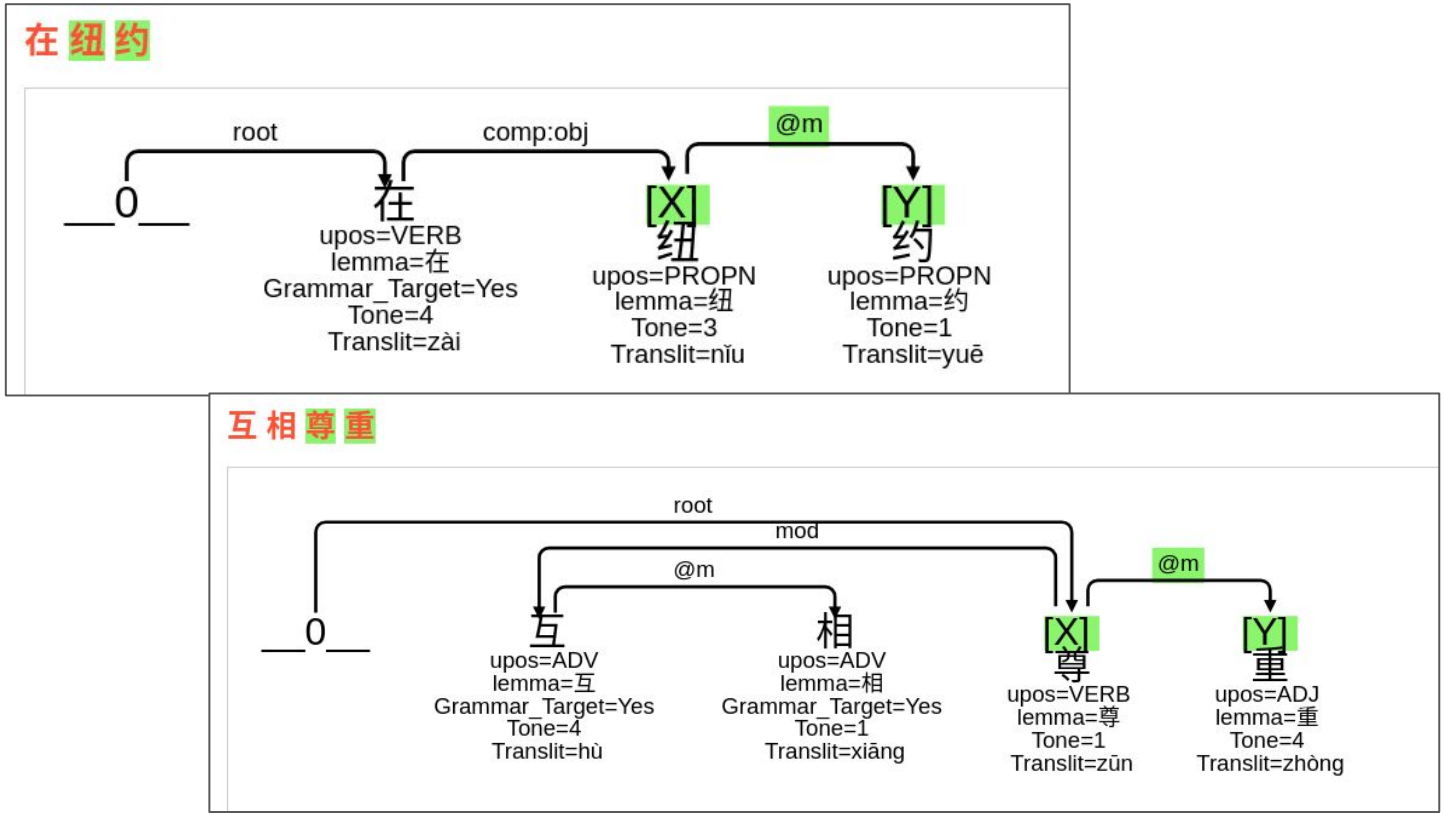
Universal dependencies :

- Basic units of annotation are syntactic words (not phonological or orthographic words)
  - dámelo [donnez-le moi] is 3 syntactic words: da me lo

Syntactic annotation at the word level can be problematic for agglutinative languages:

- Finnish: Republikaanitalousasiantuntijat [Experts économiques républicains] link
- Yupic: Mangteghaghllangllaghyugtukut [Nous voulons construire une grande maison] link

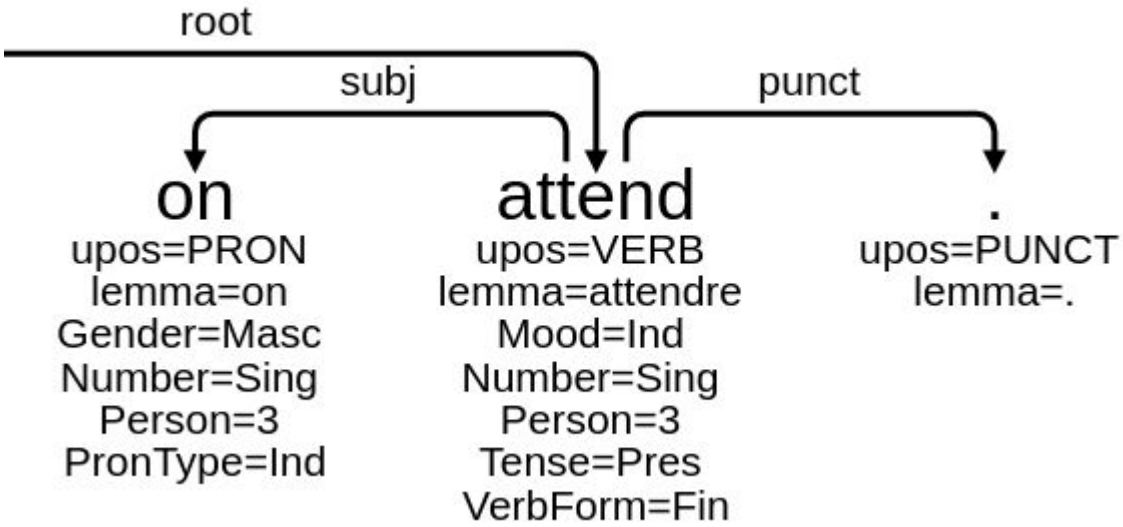
Exemple en chinois



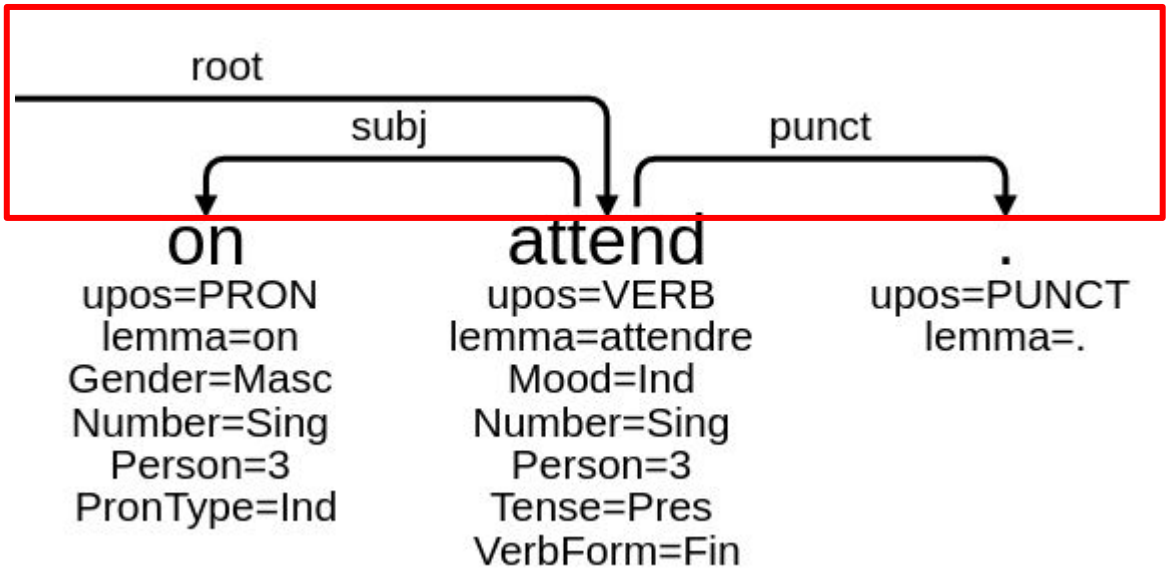
Unités d’analyse

- Première décision à prendre : segmenter en phrases
- Ensuite… Mots ? Morphèmes ? Tokens !

Couches d’annotation

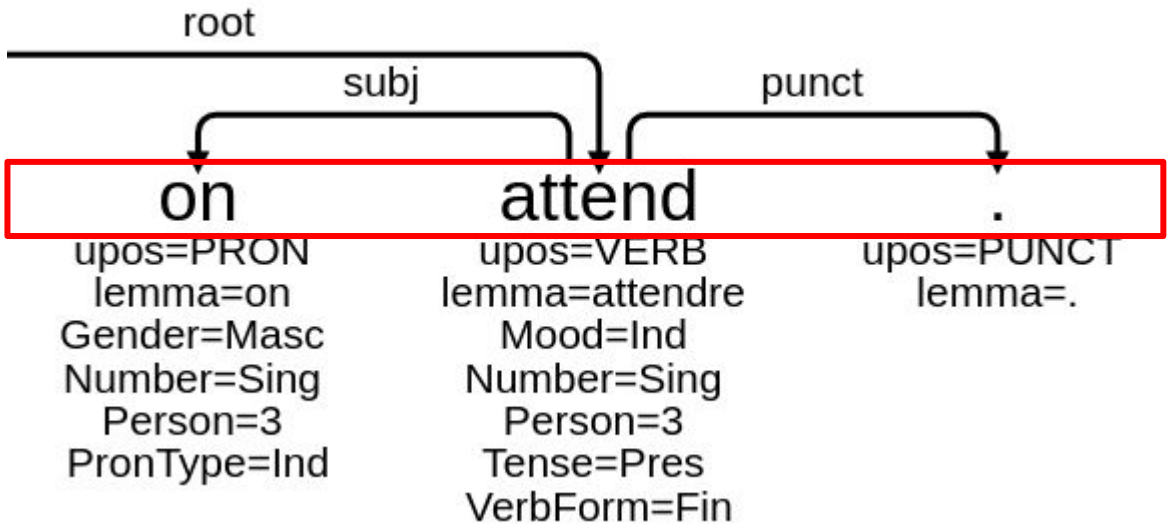


Couches d’annotation



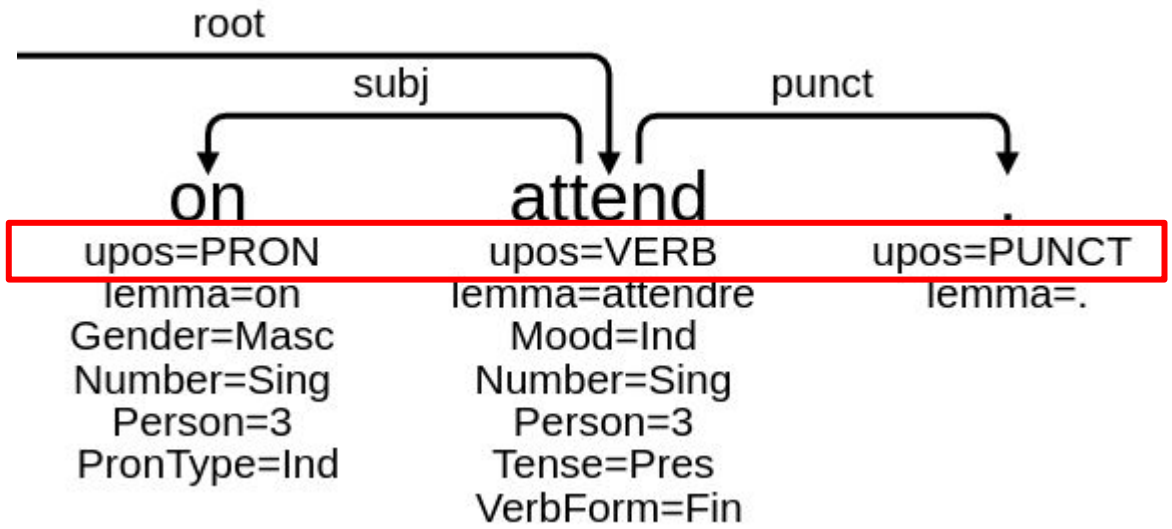
Dépendances syntaxiques

Couches d’annotation



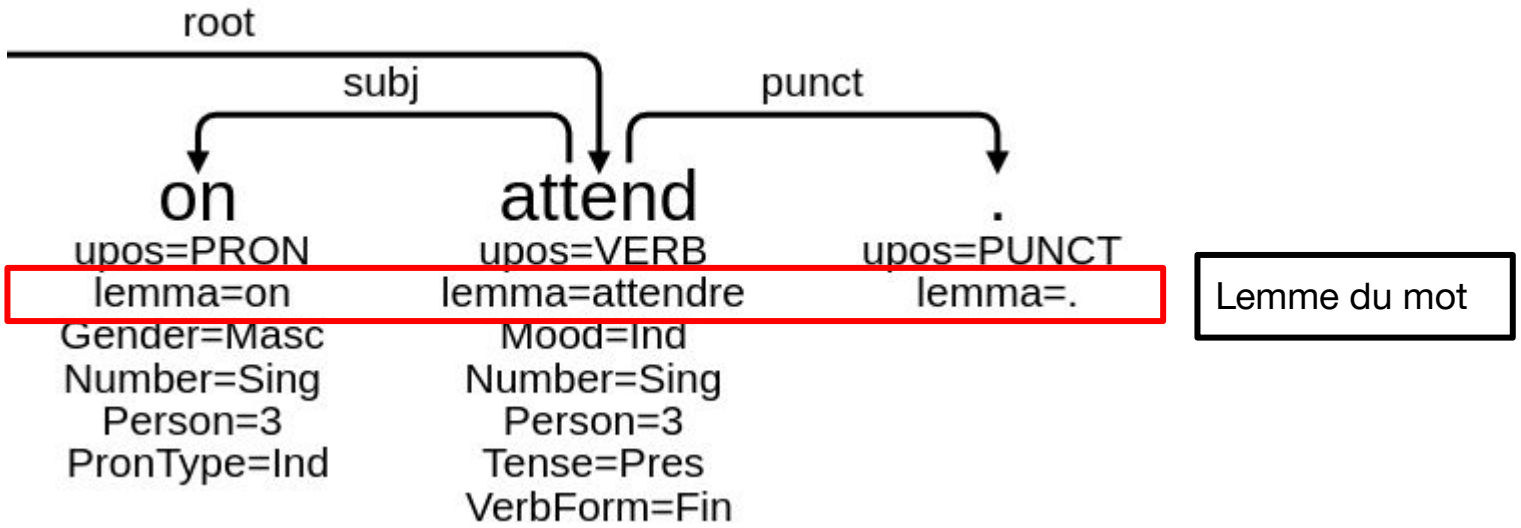
Forme textuelle / token

Couches d’annotation

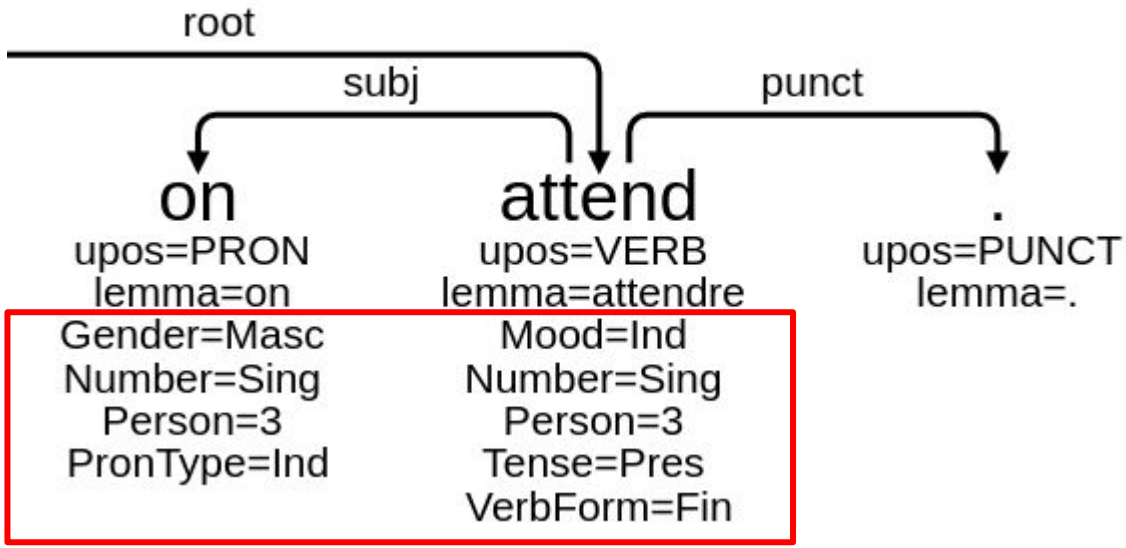


Universal Part of Speech

Couches d’annotation

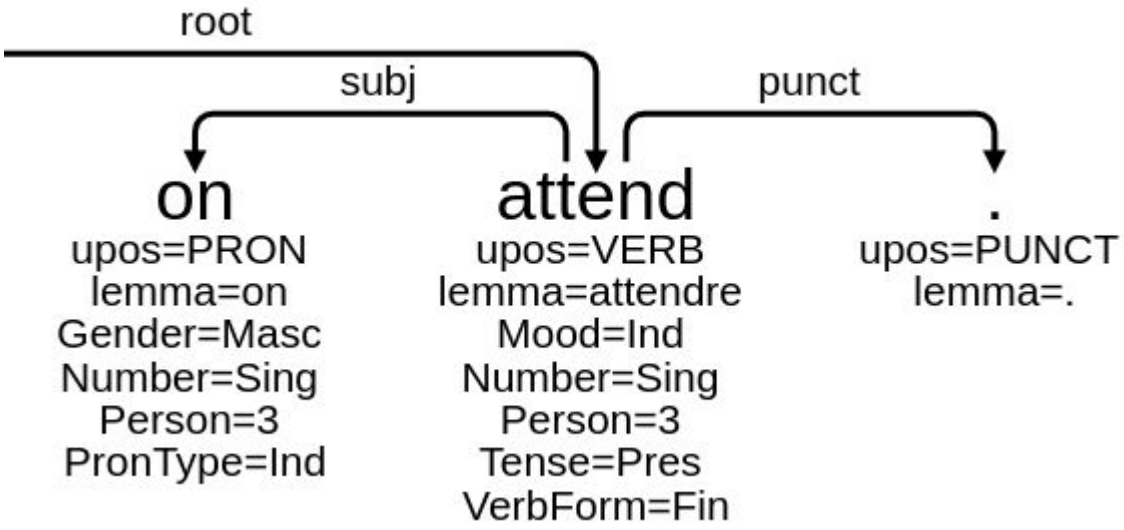


Couches d’annotation



Annotation morphosyntaxique / Traits morphologiques

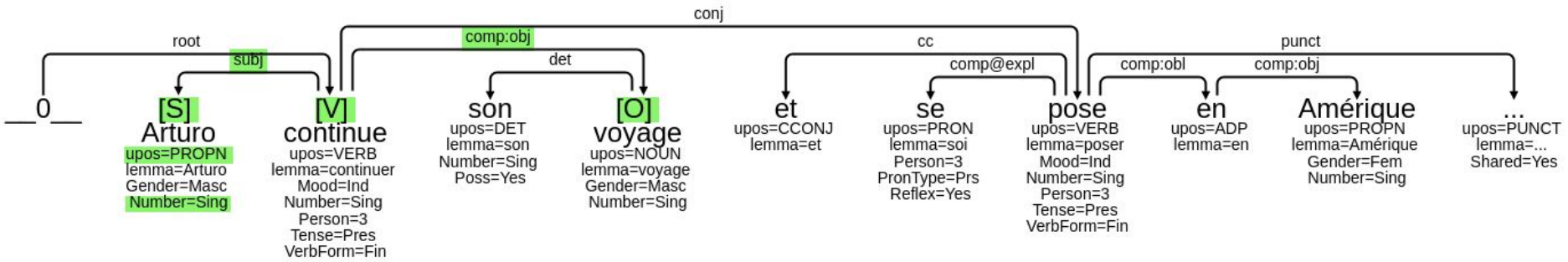
Syntaxe de Grew-match




```
pattern { X -[subj]->Y }

pattern { X [form="attend"] }

pattern { X [upos=VERB] }

pattern { X [lemma="attendre" }

pattern { X[Tense=Pres] }
```

Syntaxe de Grew-match : plusieurs contraints

```
pattern {
    V-[subj]->S;
    V-[comp:obj]->O;
    S[upos=PROPN, Number=Sing]
}
```



Syntaxe de Grew-match

```
pattern { V -[subj]-> S }
pattern { V [form="pensais"] }
pattern { V [lemma="penser"] }
pattern { V [upos=VERB, lemma="penser"]; V -[subj]-> S; V -[comp:obj]-> O }
pattern { V [upos=VERB] }
pattern { V [Tense=Past] }
pattern { V [lemma <> "manger"] }
pattern { V [upos=VERB | AUX] }
```

Le lemme est different à "see"

Le POS est soit un VERB soit un AUX

À vous

Sur le treebank GSD en SUD du français (sauf indication)

- Cherchez tous les sujets (subj)
- Trouvez tous les lemmes (lemma) des auxiliares (AUX)
- Cherchez toutes les phrases qui ont comme racine (root) un auxiliaire (AUX)
  - Réalisez la même requête en **UD**. Qu'est-ce qui se passe ?
  - Cherchez les phrases qui ont un auxiliaire **ou** un verbe comme racine
- Est-il possible d'avoir deux sujets (subj) sur le même verbe ?
- Est-il possible d'avoir deux objets (comp:obj) sur le même verbe ?

Encore Grew-match

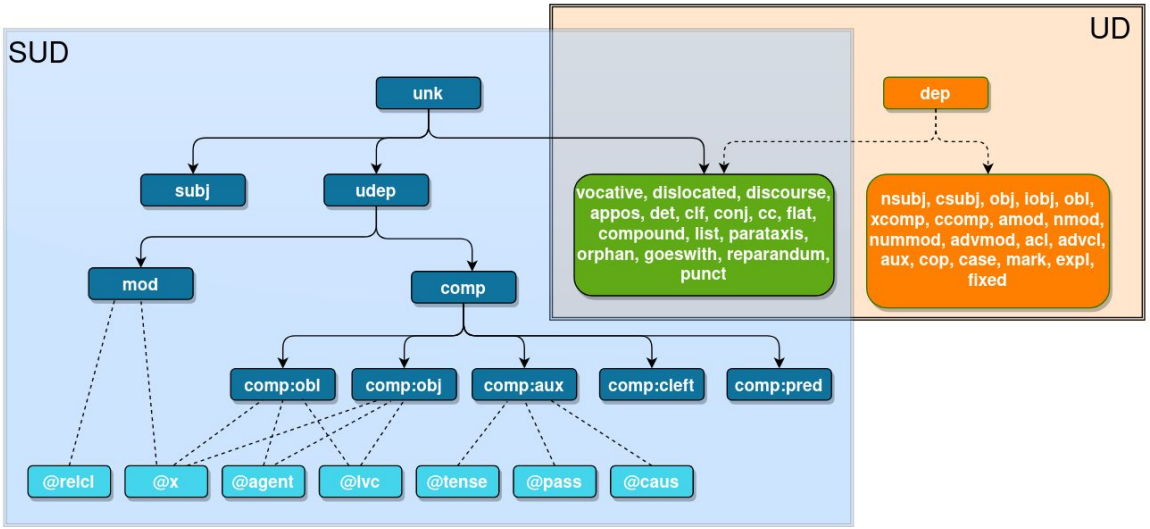
N1 < N2	Le nœud N1 est immédiatement avant N2
N1 << N2	Le nœud N1 est avant N2
N1.feature = N2.feature	Égalité entre les valeurs des features
N1.feature <> N2.feature	Inégalité entre les valeurs des features
N1[upos=Verb, !Mood]	Négation d'une feature
without { N1[upos=NOUN] }	Application de contraintes plus générales

À vous

- Y a-t-il des nœuds qui n'ont pas d'upos ?
- Y a-t-il des pairs contigus DET-NOUN qui ne soit pas dans une relation déterminative (det)
- Trouvez
  - les paires DET-NOUN contigus qui ne s'accordent pas en genre.
  - les sujets qui se placent avant le verbe.
  - les phrases verbales sans sujet
- Comment est-il annoté le trigram “à partir de” ?
- Trouvez les verbes intransitifs
- Quels sont les adjectifs sans trait de genre ?

Surface Syntactic UD

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



CONLL-U

- Qu’est-ce qui se passe avec ‘du’, ‘des’, ‘au’, ‘aux’, etc.
- Colonne MISC?

```
# 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   _   _
```

```
# sent_id = fr-ud-train_04848
# text = Gouvernement du district
1   Gouvernement    gouvernement    NOUN    _   Gender=Masc|Number=Sing 0   root    _   wordform=gouvernement
2-3 du  _   _   _   _   _   _   _   _
2   de  de  ADP _   _   1   udep   _   _
3   le  le  DET _   Definite=Def|Gender=Masc|Number=Sing|PronType=Art   4   det _   _
4   district    district    NOUN    _   Gender=Masc|Number=Sing 2   comp:obj    _   _
```