UNIVERSITÉ CÔTE D'AZUR

I.A. et Langage :
Traitement automatique du langage naturel

Elena CABRIO

elena.cabrio@univ-cotedazur.fr

Serena VILLATA

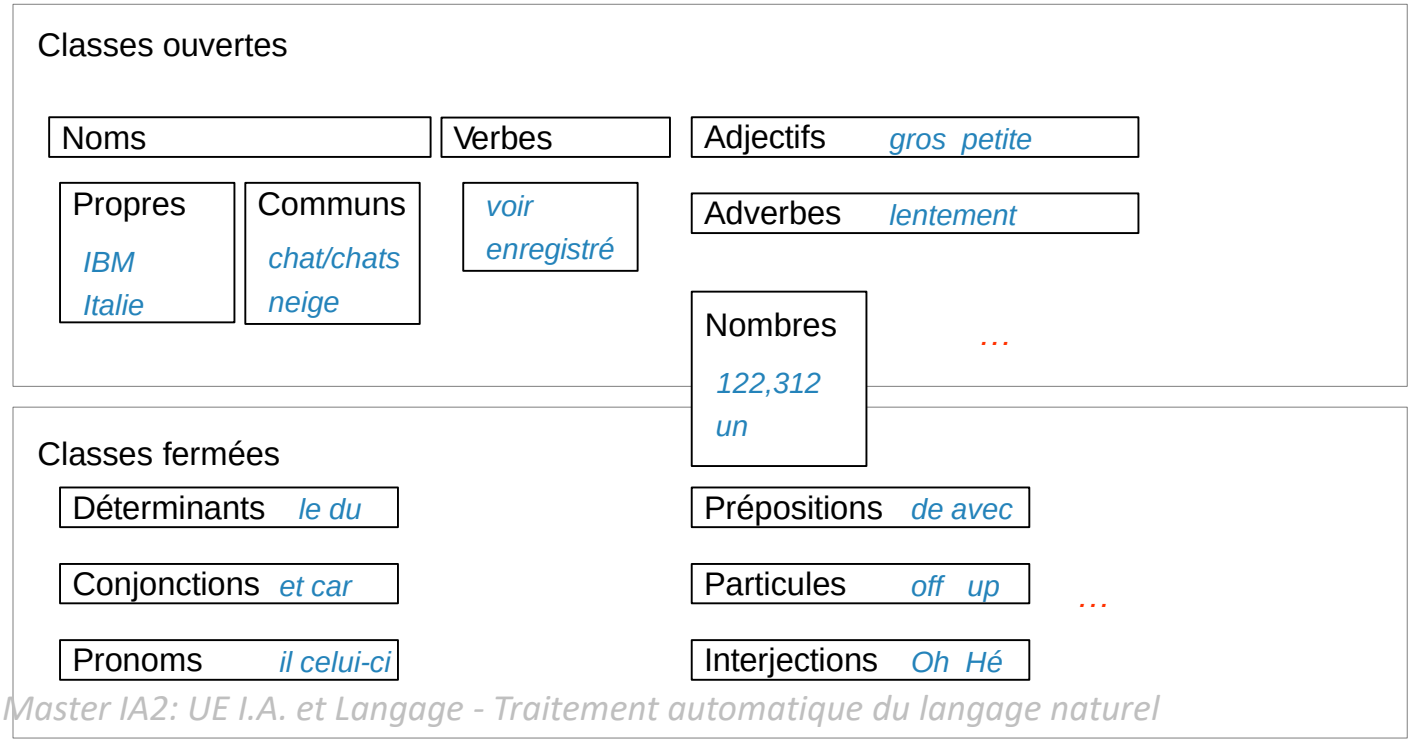
villata@i3s.unice.fr

UNIVERSITÉ CÔTE D'AZUR

Analyse morpho-syntaxique

Analyse morpho-syntaxique

BUT: analyser chaque mot pour lui associer divers types d'informations telles que **la catégorie grammaticale** (parts-of-speech), **des traits morphologiques** ainsi que le **lemme correspondant**



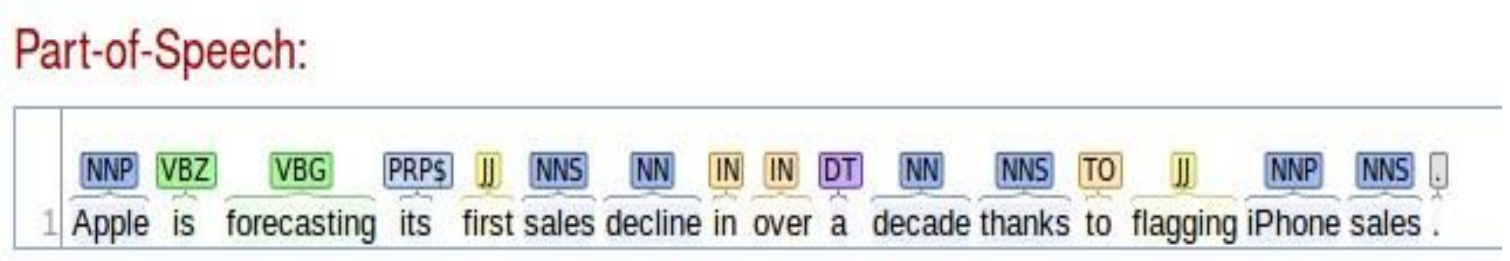
(cours C. Manning)

Etiquetage morpho-syntaxiques

UNIVERSITÉ CÔTE D'AZUR

- Les mots ont généralement plus d'une étiquette possible
  - Le bois vient de France. → le=det, bois=nom
  - Je le bois. → le = pronom, bois = verbe

- Objectif: déterminer l'étiquette pour une instance d'un mot



Exemples d'étiquetage et difficultés

UNIVERSITÉ CÔTE D'AZUR

- *Entrée:* Le débat est relancé.
  - ambiguités: le=det/pro débat=verbe/nom est=verbe/nom
- *Sortie:* Le/DET débat/NOM est/VER relancé/VER

- Applications:
  - synthèse vocale: comment prononcer *est* ?
  - recherche dans un corpus: *est* en tant que nom
  - entrée d'un analyseur syntaxique

Performance d'étiquetage

UNIVERSITÉ CÔTE D'AZUR

- Combien d'étiquettes sont correctes ? **Précision**
  - étiqueteurs sur l'anglais autour de 97%
  - mais baseline simple = 90%
    - chaque mot du lexique → étiquette la plus fréquente
    - mots inconnus → noms
  - beaucoup de mots ne sont pas ambigus
    - déterminants, prépositions, ponctuation...

Performance d'étiquetage

UNIVERSITÉ CÔTE D'AZUR

- Déterminer l'étiquette peut être difficile pour des humains également

- Un principe décliné dans la loi relative à l'informatique
- Les statistiques ethniques, c'est complètement has been
- La Commission nationale de l'informatique et des libertés (Cnil) étudie au cas par cas les demandes

Désambiguïsation des parties du discours

Elle le fait.

Elle	PRON
le	**PC**
fait	**VERB_P3SG**
.	SENT

Elle montre le fait.

Elle	PRON
montre	V_P3SG
le	**DET_SG**
fait	**NOUN_SG**
.	SENT

Sources d'information

UNIVERSITÉ CÔTE D'AZUR

- Contexte des mots

- Le bois vient de France
  - DET NOM VER PREP NAM
  - PRO VER VER PREP NAM

- Connaissance des probabilités d'étiquettes des mots

Corpus French TreeBank

- Projet initié en 1997
- http://ftb.linguist.univ-paris-diderot.fr/
- Corpus journalistique (Le Monde) 1 million de mots
- Annotations
  - Morphosyntaxique
    - POS
    - Sous-catégorisation
    - Inflection
    - Lemme
    - Parties pour mots composés
  - Constituants
  - Fonctions

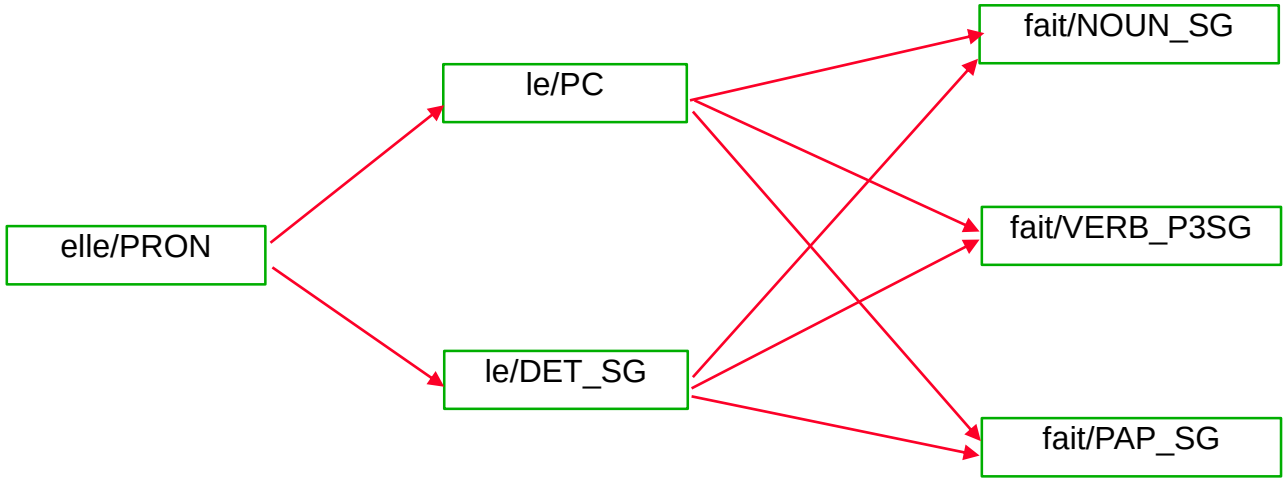
Méthodes d'étiquetage statistiques

UNIVERSITÉ CÔTE D'AZUR

- Calcul des probalitités à partir d'un corpus d'apprentissage
  - probabilités lexicales
    - prob (tag | mot) = freq (mot, tag) / freq (mot)
  - probablités contextuelles
    - ***bigrammes :***
      - prob (tag_2 | tag_1) = freq (tag_1 tag_2) / freq (tag_1)
    - ***trigrammes :***
      - prob (tag_3 | tag_1 tag_2) = freq (tag_1 tag_2 tag_3) / freq (tag_1 tag_2)

Exemple

elle le fait



Fréquences des mots et des étiquettes

UNIVERSITÉ CÔTE D'AZUR

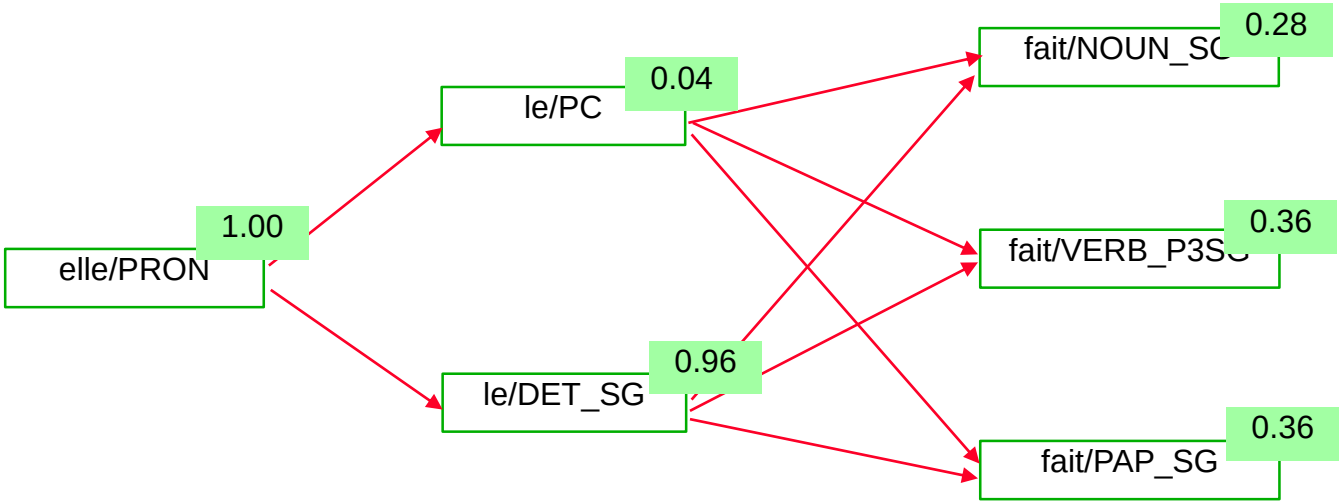
Corpus d'apprentissage: extrait "Le Monde"

freq	elle	le	fait	montre	**Tot.**
PRON	17	--	--	--	**320**
DET_SG	--	239	--	--	**1329**
PC	--	11	--	--	**179**
VERB_P3SG	--	--	5	2	**371**
NOUN_SG	--	--	4	0	**1931**
PAP_SG	--	--	5	--	**207**
...	...	…	...	...	**...**
Tot.	**17**	**250**	**14**	**2**	**15.000**

Calcul des probabilités lexicales

prob (PRON | elle)	= 17 / 17 = 1.00
prob (DET_SG | le)	= 239 / 250 = 0.96
prob (PC | le)	= 11 / 250 = 0.04
prob (NOUN_SG | fait)	= 4 / 14= 0.28
prob (PAP_SG | fait)	= 5 / 14 = 0.36
prob (VERB_P3SG | fait)	= 5 / 14 = 0.36

Probabilités lexicales



prob (elle/PRON le/PC fait/VERB_P3SG) = 1.00 * 0.04 * 0.36 = 0.014

prob (elle/PRON le/DET_SG fait/NOUN_SG) = 1.00 * 0.96 * 0.28 = 0.269

prob (elle/PRON le/DET_SG fait/VERB_P3SG) = 1.00 * 0.96 * 0.36 = 0.346

Fréquences des séquences d'étiquettes

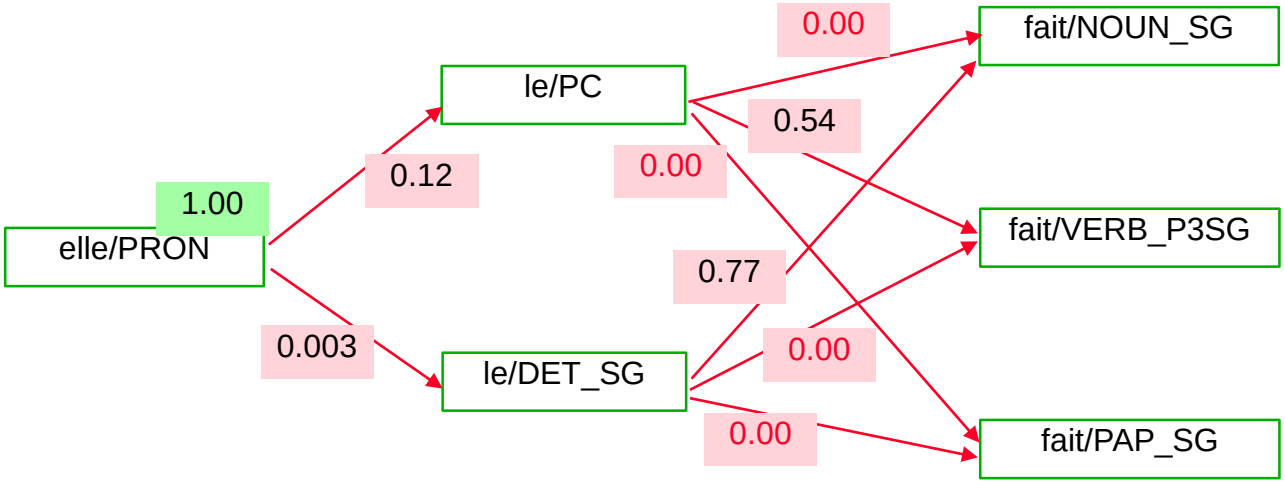
tag_1 \ tag_2	PRON	DET_SG	PC	VERB_P3SG	N_SG	PAP_SG	**Tot.**
PRON	--	1	38	82	--	32	**320**
DET_SG	4	5	--	--	1033	--	**1329**
PC	--	--	3	59	--	--	**179**
VERB_P3SG	17	53	10	--	9	--	**371**
NOUN_SG	3	29	12	46	13	1	**1931**
PAP_SG	1	42	--	1	10	--	**207**
...	...	...	...	...	...	...	**...**
Tot.	**320**	**1329**	**179**	**564**	**1931**	**207**	**15.000**
				...			

Calcul des probabilités contextuelles

UNIVERSITÉ CÔTE D'AZUR

prob (PC | PRON)	= 38 / 320 = 0.12
prob (DET_SG | PRON)	= 1 / 320 = 0.003
prob (VERB_P3SG | PC)	= 97 / 179 = 0.54
prob (PAP_SG | PC)	= 0 / 179 = 0.00
prob (NOUN_SG | PC)	= 0 / 179 = 0.00
prob (NOUN_SG | DET_SG)	= 1033 / 1329 = 0.77
prob (VERB_P3SG | DET_SG)	= 0 / 1329 = 0.00
prob (PAP_SG | DET_SG)	= 0 / 1329 = 0.00
...	

Probabilités contextuelles

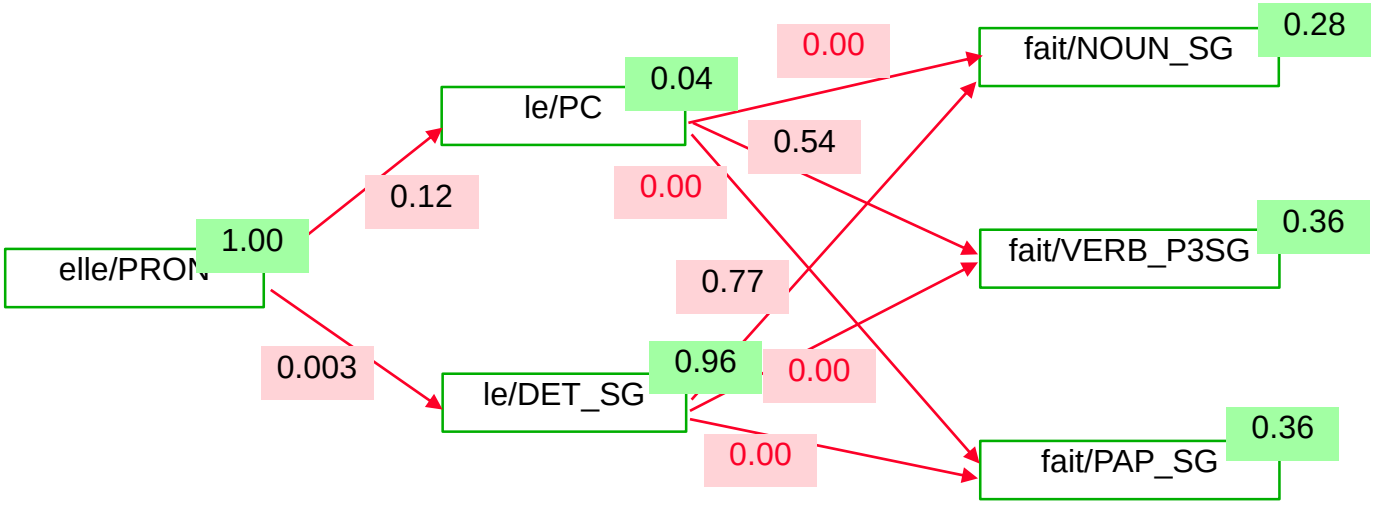


prob (elle/PRON le/PC fait/VERB_P3SG) = 0.12 * 0.54 = 0.0648

prob (elle/PRON le/DET_SG fait/NOUN_SG) = 0.003 * 0.77 = 0.0231

prob (elle/PRON le/DET_SG fait/VERB_P3SG) = 0.003 * 0.00 = 0

Probabilités lexicales et contextuelles



prob (elle/PRON le/PC fait/VERB_P3SG) = 1.00 * 0.12 * 0.04 * 0.54 * 0.36 = 0.00093

prob (elle/PRON le/DET_SG fait/NOUN_SG) = 1.00 * 0.003 * 0.96 * 0.77 * 0.28 = 0.00058

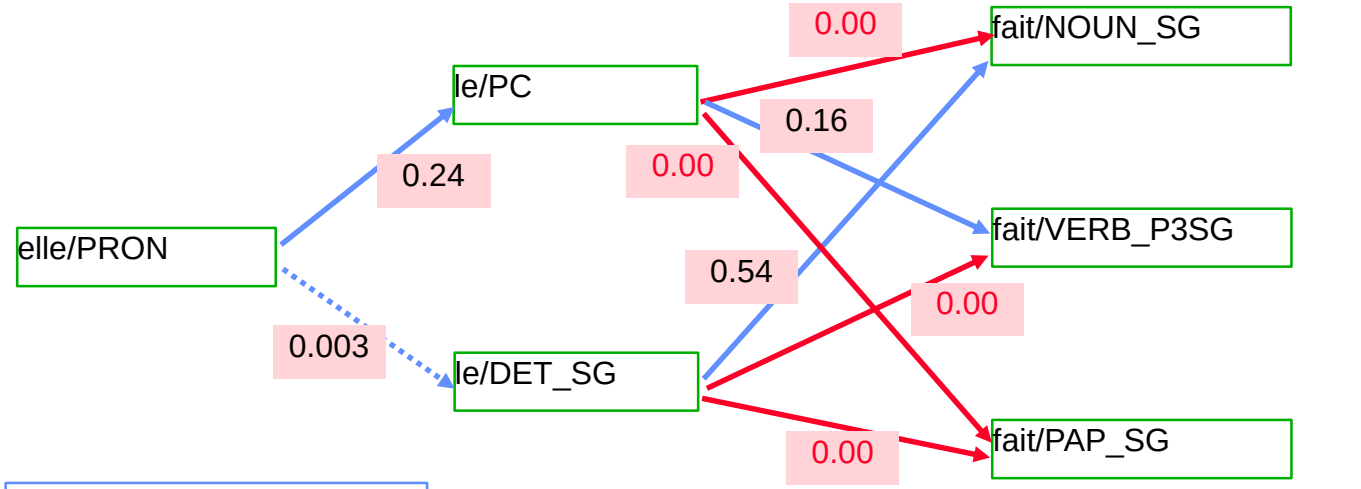
prob (elle/PRON le/DET_SG fait/VERB_P3SG) = 1.00 * 0.003 * 0.96 * 0.00 * 0.36 = 0

Méthodes à base de règles

UNIVERSITÉ CÔTE D'AZUR

- règles “positives”
  - pour définir les séquences possibles
  - exemple:
    - un pronom personnel est suivi d’un verbe
    - un déterminant est suivi d’un nom
- règles “négatives”
  - pour exclure des séquences impossibles
  - exemple:
    - un pronom enclitique ne précède pas un nom
    - un déterminant ne précède pas un verbe

Probabilités et règles



PRON => PC

PC => VERB_P3SG

DET_SG => NOUN_SG

PC ≠>NOUN_SG

PC ≠> PAP_SG

DET_SG ≠> VERB_P3SG

DET_SG ≠> PAP_SG

Méthodes à base de « transformations »

UNIVERSITÉ CÔTE D'AZUR

- Brill tagger.
- L'idée générale très simple: deviner l'étiquette de chaque mot, puis revenir en arrière et corriger les erreurs. De cette façon, un tagger Brill transforme successivement un mauvais marquage d'un texte en un meilleur (**méthode d'apprentissage supervisée**).
- Contrairement au marquage n-gram, il ne compte pas les observations mais compile une liste de règles de ``correction transformationnelle''.
- Les règles sont linguistiquement interprétables
- https://www.nltk.org/api/nltk.tag.html#module-nltk.tag.brill

Modèles de Markov cachés

UNIVERSITÉ CÔTE D'AZUR

- Les modèles de Markov cachés (HMM) sont largement utilisés pour attribuer la séquence d'étiquettes correcte à des données séquentielles ou pour évaluer la probabilité d'une étiquette et d'une séquence de données données.

- Ces modèles sont des machines à états finis caractérisés par un certain nombre d'**états**, des **transitions entre ces états** et des **symboles de sortie** émis dans chaque état.

- Le HMM est une extension de la chaîne de Markov, où chaque état correspond de manière déterministe à un événement donné. Dans le HMM, l'observation est une fonction probabiliste de l'état.

- Les HMM partagent l'hypothèse de la chaîne de Markov, à savoir que la probabilité de transition d'un état à un autre ne dépend que de l'état actuel - c'est-à-dire que la série d'états ayant conduit à l'état actuel n'est pas utilisée.

Modèles de Markov cachés

- Le HMM est un graphe orienté, avec des arêtes pondérées en fonction de la probabilité (représentant la probabilité d'une transition entre les états source et récepteur), où chaque sommet émet un symbole de sortie lorsqu'il est entré. Le symbole (ou observation) est généré de manière non déterministe.
- Pour cette raison, le fait de savoir qu'une séquence d'observations en sortie a été générée par un HMM donné ne signifie pas que la séquence d'états correspondante (et ce qu'est l'état actuel) est connue. C'est le "caché" dans le modèle de Markov caché.
- Un HMM est souhaitable pour la tâche d' étiquetage morpho-syntaxique car la séquence d'étiquettes présentant la plus grande probabilité peut être calculée pour une séquence donnée de mots. Pour tenir compte de la combinaison optimale des tags pour une unité plus grande, telle qu'une phrase, le HMM exploite l'algorithme de Viterbi, qui calcule efficacement le chemin optimal à travers le graphe étant donné la séquence de mots.
- https://www.nltk.org/_modules/nltk/tag/hmm.html

Natural language toolkit (+ Français)

UNIVERSITÉ CÔTE D'AZUR

Les étiqueteurs grammaticaux sont très nombreux pour les langues saxonnes mais plus rares pour le français. Des étiqueteurs sont accessibles avec un modèle pour le français prêt à l'emploi, des autres peuvent fonctionner pour le français mais doivent être entraînés sur un corpus français pré-étiqueté.

NLTK (Natural Language Toolkit)
http://www.nltk.org/

Stanford Parser et CoreNLP (méthodes statistiques**)**
https://stanfordnlp.github.io/CoreNLP/
http://corenlp.run/
http://nlp.stanford.edu:8080/parser/index.jsp

TreeTagger
https://cis.uni-muenchen.de/~schmid/tools/TreeTagger/

spaCy (méthodes deep**)**
https://spacy.io/
https://explosion.ai/demos/

UNIVERSITÉ CÔTE D'AZUR

Traitement automatique de base

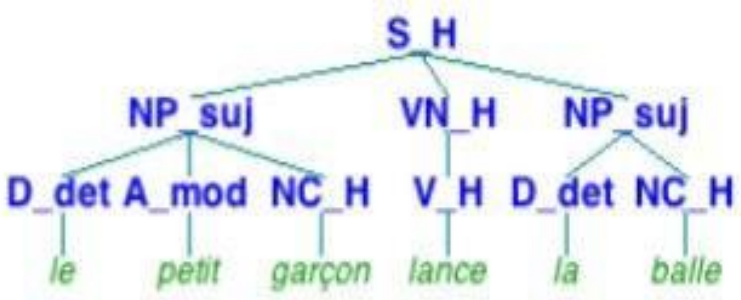
Analyse syntaxique

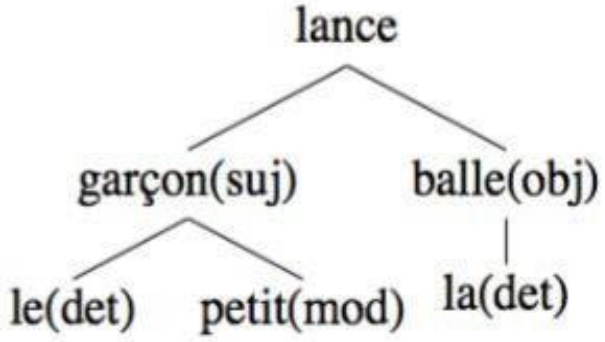
Objectif de l'analyse syntaxique

UNIVERSITÉ CÔTE D'AZUR

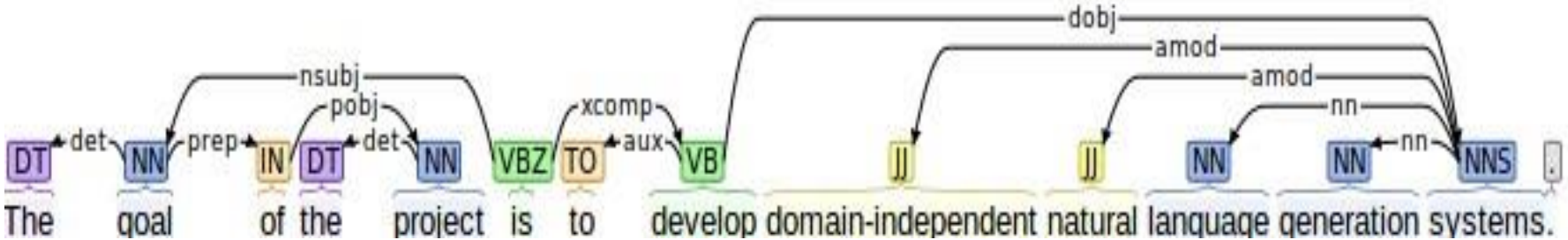
- Analyse syntaxique traditionnelle
  - Généralement fondée sur le paradigme génératif de Chomsky
  - Objet = générer tous et seulement les énoncés possibles dans une langue (énoncés grammaticaux)
  - En analyse = associer à un énoncé (phrase) grammatical(e) de la langue sa structure syntaxique
    - arbre des séquences de réécritures permettant d'obtenir la phrase à partir de l'axiome S de la grammaire

Exemple de sortie attendue





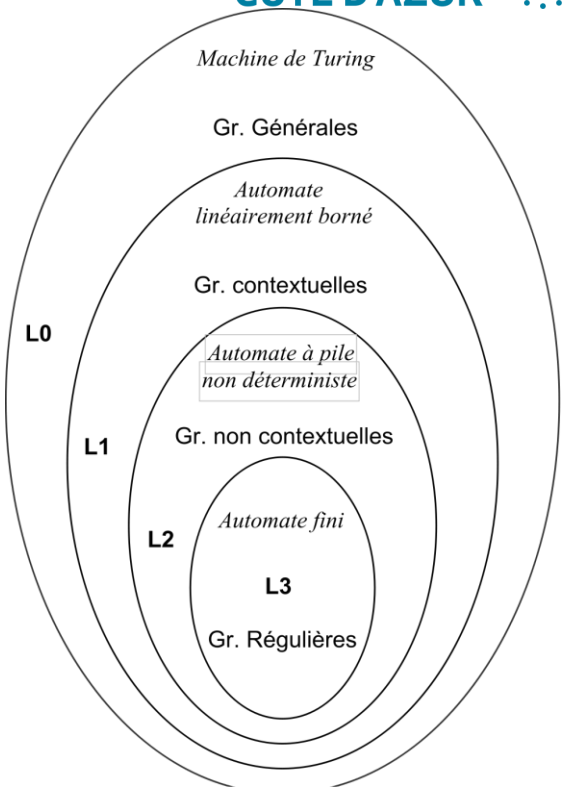
Exemple d'analyse en dépendances



Grammaires

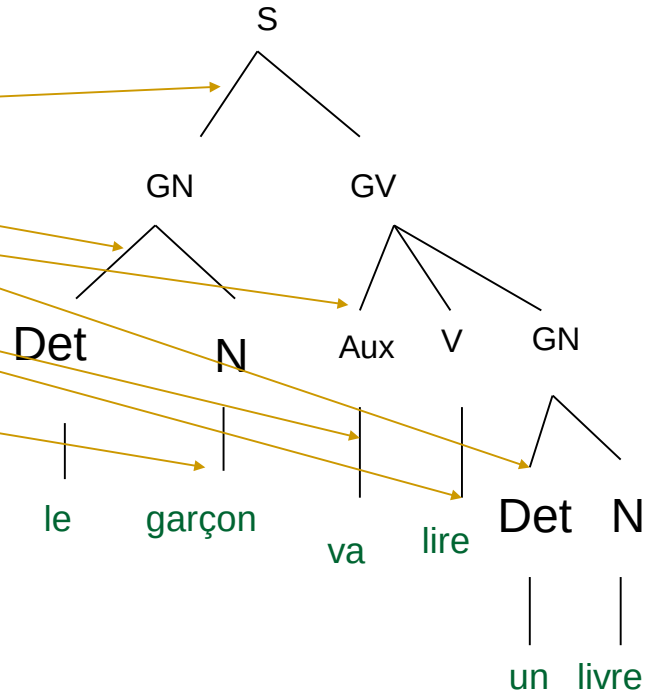
- G=(Vn, Vt, R, S)
  - Vn : vocabulaire non terminal
  - Vt : vocabulaire terminal
  - R : ensemble de règles de réécriture
  - X → Y S : axiome de la grammaire

- Suivant les règles de R
  - Grammaire non contrainte → trop « lâche »
  - Grammaire en contexte :
    - « X se réécrit Y dans le contexte u v »
    - uXv → uYv
  - Grammaire hors contexte : X → Y
  - Grammaire régulière (trop figée)
    - A → a ou A → aB



Grammaires hors-contexte

- Exemple :
- S → GN GV
- GN → Det N
- GV → (Aux) V GN
- Aux → va
- V→lire | bat |mange|…
- Det → le|la|les|un…
- N → garçon|livre|pomme…

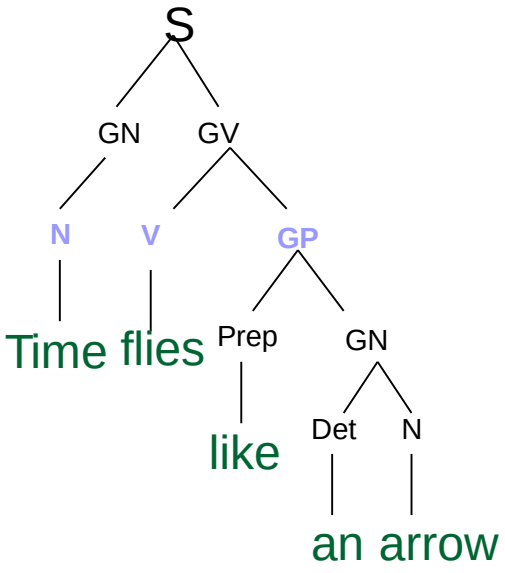


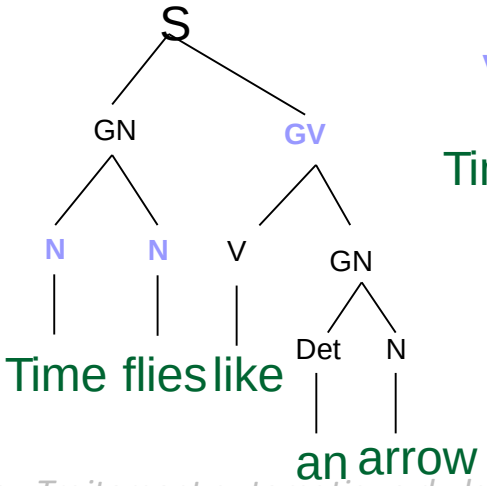
Le garçon va lire un livre

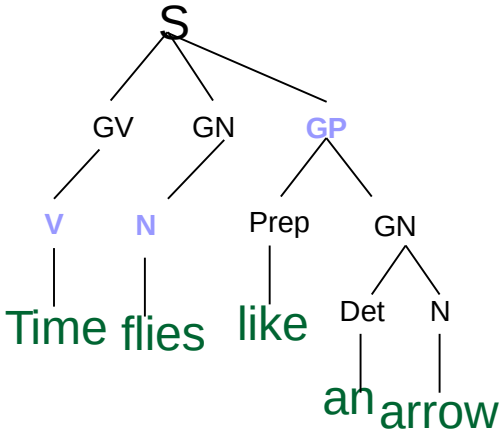
Mais aussi : *le pomme va mange la livre*

Grammaires hors-contexte

- Différences entre structure de surface et structures profondes
- Exemple « chomskyen » : Time flies like an arrow:







Analyse syntaxique traditionnelle

UNIVERSITÉ CÔTE D'AZUR

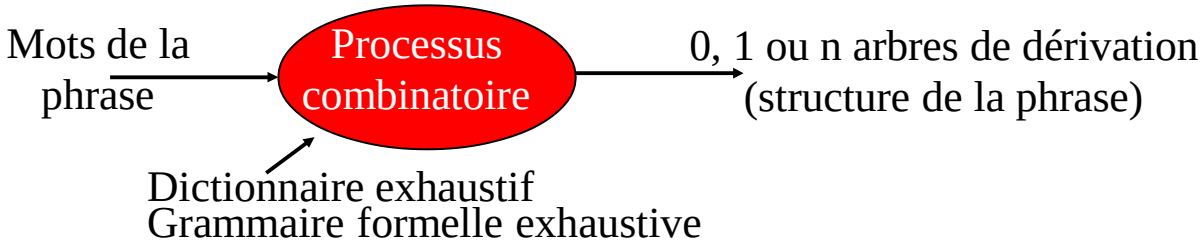
- Théorie des langages formels de Chomsky
  - Formalisation mathématique pas une théorie linguistique
  - La langue n'est pas un langage indépendant du contexte
    - Les accords
  - Grammaires contextuelles insuffisantes
    - Constituants discontinus : Combien cette salle a-t-elle de fenêtres ?

Examples d'analyseurs: DCG (Definite Clause Grammar)

```
sentence --> noun_phrase, verb_phrase.
noun_phrase --> det, noun.
verb_phrase --> verb, noun_phrase.
det --> [le].
noun --> [chat].
noun --> [chauve-souris].
verb --> [mange]
```

- Cela génère des phrases telles que "le chat mange la chauve-souris", "une chauve-souris mange le chat".
- On peut vérifier si une phrase est valide dans la langue en tapant quelque chose comme sentence ([le, chat, mange, le, chauve-souris], []) (Prolog)

Analyse syntaxique traditionnelle

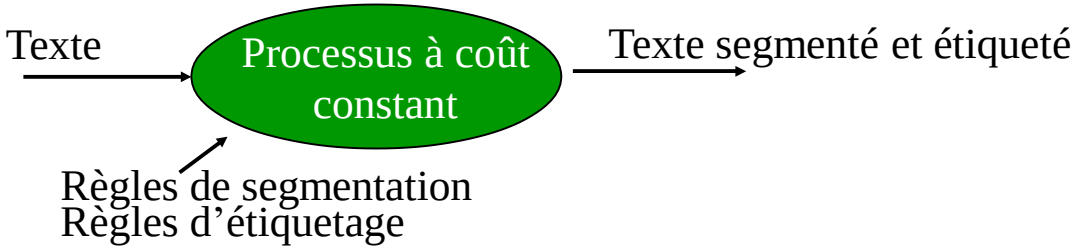


- Caractéristiques (HPSG, LFG, TAG, …) :
  - Règles de grammaire de type hors-contexte
  - Structures de traits
  - Unification
- Problème : manque de robustesse

Analyse robuste

UNIVERSITÉ CÔTE D'AZUR

Analyse robuste, analyse partielle, analyse de surface (shallow parsing)



- Approche empirique : héritage de la reconnaissance de la parole
- Travail sur texte réel, but opérationnel d'abord
- Analyse vue comme un processus informatique
- Principalement des méthodes statistiques

Notion de robustesse en TAL

UNIVERSITÉ CÔTE D'AZUR

- Robustesse : plusieurs définitions dans la littérature du TAL
- Idée commune :
  - Capacité d'un système de TAL à traiter des données linguistiques réelles (produites par des locuteurs indépendamment du système)

- Définition (pour un analyseur)
  - Capacité d'un système à produire des analyses utiles pour des textes réels
  - Analyses utiles : analyses (au moins partiellement) correctes et utilisables dans une tâche automatique (application)

Propriétés nécessaires

UNIVERSITÉ CÔTE D'AZUR

- Une analyse au moins pour chaque entrée
  - Situations d'absence d'analyses fréquentes dans les analyseurs traditionnels
  - Enoncés agrammaticaux dans les textes réels
  - Mais, plus fréquemment : constructions grammaticales non prédites par le modèle ou les descriptions linguistiques de l'analyseur

- Nombre d'analyses concurrentes limité
  - Les analyseurs traditionnels produisent souvent de trop nombreuses analyses (parfois des milliers pour une longue phrase), dont des analyses redondantes (ambiguïtés artificielles)

Méthodes d'analyse robuste

UNIVERSITÉ CÔTE D'AZUR

- Emergence de méthodes d'analyse robuste

- Trois tendances générales
  - Ajout de mécanismes ad hoc spécifiques pour rendre les analyseurs traditionnels robustes
  - Analyse à base de modèles statistiques
  - Analyse de surface à base de règles (rule-based *shallow parsing*)

Analyse de surface (*shallow parsing*)

UNIVERSITÉ CÔTE D'AZUR

- Idée de base
  - Limiter la « profondeur » et la richesse de l'analyse syntaxique
  - Prévoir la possibilité d'analyses partielles

- But
  - Obtenir des structures syntaxiques minimales, sous- spécifiées mais linguistiquement motivées (syntagme noyau = *chunk*)
  - Des structures utiles en tant que telles dans des applications
  - Première phase d'une analyse syntaxique plus complète

Exemple d'analyse

- [Bill NP] [vit V] [l'homme NP] [sur la colline PP] [avec un télescope PP]

- Chunks : NP, V, PP
- Ambiguïté de rattachement implicite

Analyse de surface: étapes de traitements

- Prétraitement
  - Etiquetage morpho-syntaxique (segmentation, analyse morphologique, désambiguïsation)
- Analyse syntaxique de surface
  - Reconnaissance des syntagmes noyaux (chunks) : SN, SP, SV
  - Groupes complexes et propositions
  - Attribution de fonctions syntaxiques (Sujet, Objet, etc.)
- Analyse incrémentale

Analyse par apprentissage supervisé

UNIVERSITÉ CÔTE D'AZUR

- Nécessité de grands corpus annotés
  - Penn TreeBank pour l'anglais
  - French TreeBank pour le français

Représentation dans le FTB