Introduction aux principes de traitement de données textuelles

François Bouchet
LIP6 / SU
francois.bouchet@lip6.fr

28 septembre 2018

SORBONNE UNIVERSITÉ

http://bit.ly/EVIJV-2018-1

LIP6

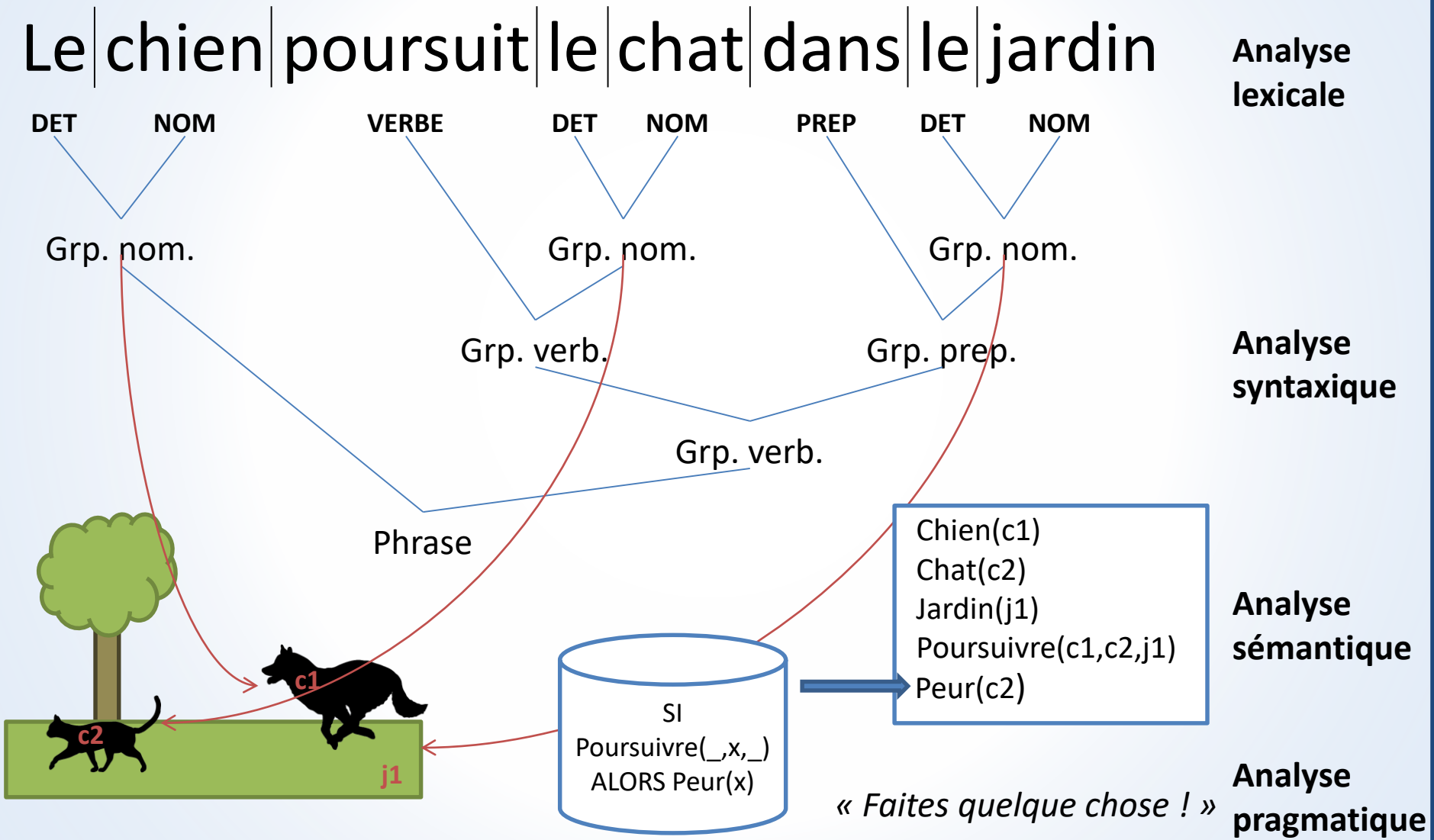
Plan du cours

1. **Introduction**
   - Utilité du traitement de la langue naturelle
   - Principes de base du traitement de la langue naturelle
   - Rappels historiques
   - Types de problèmes
2. **Analyse lexicale**
3. **Analyse syntaxique**
4. **Analyse sémantique**

Intérêt du traitement automatique de la langue naturelle

- Classification de textes
- Système de question-réponse (e.g. moteur de recherche sémantique)
- Dialogue homme-machine
- Traduction automatique
- Analyse de productions textuelles (e.g. résumés)
- Génération automatique de questions associées à un contenu
- Analyse de sentiments
- Système d'aide intégré (questions-réponses)
- …

Concepts de base



Principes de base du traitement de la langue naturelle

- Différents niveaux d'analyse :
  - Phonétique (sons)
  - Morphologie (composants des mots)
  - Syntaxe et grammaire (agencement des mots)
  - Sémantique (sens des mots et de la phrase)
  - Pragmatique (sens par rapport aux intentions)
  - Discours (agencement de phrases)
- Pas forcément séquentiels
- Ambiguïtés à tous les niveaux !

Différents types de problèmes

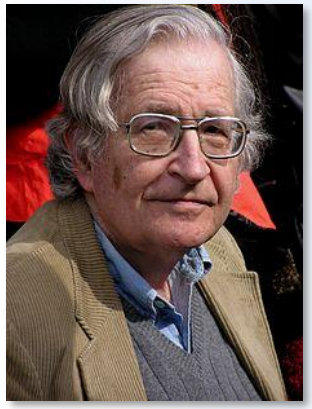
- Au niveau des **mots** :
  - Ambiguïtés de formes : nourrissons (nous + nourrir au présent / nourrisson pluriel)
  - Formes variables, même concept : clé / clef
  - Formes différentes, même concept (synonymes) : soulier / chaussure
  - Une forme, des concepts différents (polysémie) : permis (autorisation) / permis (papier)

- Au niveau de la **phrase** :
  - Ambiguïté de forme :
    il regarde l'aile de l'avion (appartenant à l'avion / depuis un siège de l'avion)
  - Formes variables, même sens (allotaxie et paraphrases) :
    le drapeau s'agite au vent / le vent agite le drapeau
  - Formes identiques, sens différents (homotaxie) :
    prendre son train / prendre son parapluie
  - Métonymies : perdre la tête (contenant pour contenu)
  - Métaphores : ce vieillard est une tortue
  - Présuppositions : « il a allumé la lumière », donc elle était éteinte

- Au niveau du **discours** :
  - Anaphores : Jean prit son stylo. Il écrivait bien.
  - Ellipses : Que manges-tu ? Une pomme.

Une brève histoire du traitement automatique des langues

- Années 1940-50 : travaux fondateurs
  - Automates, grammaires hors contexte [Chomsky] (textes – innéité, pauvreté du stimulus)
  - Modèles probabilistes [Shannon] (acoustique)
  - Travaux centrés sur traduction automatique (guerre froide)
  - 1956 : création du terme IA



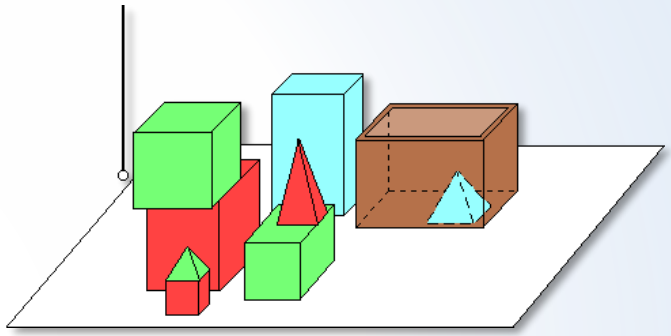
Noam Chomsky

- Années 60 : premiers systèmes
  - Algorithmes d'analyse grammaticale
  - 1966 : ELIZA, premier chatbot
  - 1966 : rapport ALPAC (« the pen is in the box » / « the box is in the pen » - [Bar-Hillel])

John R. Pierce, ALPAC

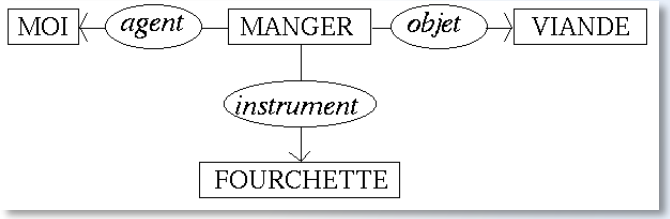
Une brève histoire du traitement automatique des langues (2)

- Années 70-80 : progrès de la sémantique formelle
  - 1972 : SHRDLU (interaction langagière, monde de blocs)
  - Nouvelles logiques (floues, modales…), scripts [Schank], frames [Minksky], graphes conceptuels [Sowa], systèmes experts (même modèles symboliques)



Monde de cubes

- Depuis les années 90 : linguistique de corpus
  - Explosion de la capacité de stockage
  - Approches statistiques
  - "Every time I fire a linguist, the performance of the speech recognizer goes up“ [Jelinek – reconnaissance vocale]



Graphe conceptuel

Plan du cours

1. Introduction

2. **Analyse lexicale**
   - Expressions régulières & automates à états finis
   - Prétraitement
   - Principes
   - Étiquetage morpho-syntaxique (TreeTagger)

3. Analyse syntaxique

4. Analyse sémantique

Expressions régulières (1)

- Développées dès l'aube du TALN [Kleene, 1956]

- **Expression régulière** (ER, RE, regex) : formule dans un langage spécial (standard) permettant d'exprimer des chaînes de caractères
- Chaîne de caractères : séquence de symboles alphanumériques

- Objectifs :
  - Rechercher des chaînes de caractères particulières dans une phrase ou un document
  - Remplacer une chaîne de caractère par une autre

Expressions régulières (2)

- **Disjonction de caractères** : [$c_1c_2c_3$]
  - /[tT]est/ matche « test » ou « Test »
  - /[1234567890]/ matche n'importe quel chiffre
- **Ensemble de caractères** : [c_1-c_2]
  - /[a-z]/ matche tout caractère en minuscule
  - /[0-9]/ matche n'importe quel chiffre (unique)
- **Négation** : [^c_1]
  - /[^A-Z]/ matche tout caractère qui n'est pas une majuscule
  - /[^tT]/ matche tout caractère qui n'est pas t ou T
- **Répétition 0/1 à n fois** : c_1*, c_1+, c_1?
  - /ba*/ matche une séquence de b suivi de 0 ou plus a (b, ba, baa…)
  - /[0-9]+/ matche une séquence de 1 ou plusieurs chiffres
  - /clef?/ matche « cle » ou « clef » mais pas « cleff »
- **N'importe quel caractère** : .
  - /d.but/ matche début, debut, ddbut, dabut…
  - /test.*/ matche n'importe quel chaîne de caractère commençant par test

Expressions régulières (3)

- **Limite de mot** : \b
  - /\bla\b/ matche « la » mais pas « larme » ou « maladresse »
- **Début/fin de ligne** : ^ / $
  - /^Le chat$/ matche une ligne contenant uniquement « Le chat » mais pas « Le chat dort »
- **Disjonction** : |
  - /chat|chien/ matche le mot « chat » ou « chien »
- **Précédence** : (et)
  - /chien(ne|)(s|)/ matche « chien », « chiens », « chienne » ou « chiennes »
- **Substitution** : s/regexp/pattern
  - s/clef/clé remplace « clef » par « clé »
  - s/rouge|jaune|vert/\1s/ remplace « rouge », « jaune » ou « vert » par leur pluriel

Eliza

- Eliza [Weizenbaum, 1966] est une simulation d'un psychanalyste et l'ancêtre des agents conversationnels ou chatterbots

- Exemple d'interaction :
  - USER: Men are all alike
  - ELIZA: IN WHAT WAY?
  - USER: They're always bugging us with one thing or another
  - ELIZA: CAN YOU THINK OF A SPECIFIC EXAMPLE?
  - USER: My boyfriend made me come here
  - ELIZA: YOUR BOYFRIEND MADE YOU COME HERE
  - USER: He says I'm depressed much of the time
  - ELIZA: I AM SORRY TO HEAR YOU ARE DEPRESSED?

Substitution d'expressions régulières

```
s/.* I AM (depressed|sad) .*/
I AM SORRY TO HEAR YOU ARE \1?/

s/.* I AM (depressed|sad) .* /
WHY DO YOU THINK YOU ARE \1?/

s/.* all .*/ IN WHAT WAY?

s/.* always .* /
CAN YOU THINK OF A SPECIFIC
EXAMPLE?
```

Prétraitements

- Segmentation : comment découper ?
  - , et . = séparations de propositions
    - M. Jacques réclame 58,50 euros à la S.N.C.F
  - - = séparateur de mots composés
    - Les Bleus ont gagné 3-0
  - ’ = élision
    - Mme D’Arcy dit : « j’viens aujourd’hui »
- Format :
  - Encodage : ASCII, UTF-8…
  - Texte brut vs. enrichi (RTF, HTML, etc.)

Analyse lexicale (1)

- Objectif : passage de formes atomiques (tokens) à des mots

- Qu'est-ce qu'un mot ?
  - Aujourd'hui (au jour d'hui)
  - Ceci vs. ceux-ci ?
  - Pomme de terre
  - Au (= à + le)

- Unité linguistique dotée de caractéristiques propres (sens, prononciation, propriétés syntaxiques…)

Analyse lexicale (2)

- Lexique pré-compilé de la langue mais :
  - Domaines spécialisés
  - Noms propres (reconnaissance d'entités nommées) :
    - « le Prix Goncourt » : récompense, montant financier, personne, jury, objet, événement…
  - Création constante de vocabulaire :
    - Macronisme, macronphobe, macronien…
- Détour nécessaire par la morphologie :
  - Découpage des mots en morphèmes :

    Anti-constitu-tion-nelle-ment
  - Ajustement de forme lié aux conditions syntaxiques (flexions) : -s pour le pluriel, -e pour le féminin d'un adjectif…
  - Créations de nouvelles formes

Lemmatisation & Étiquetage morpho-syntaxique

- **Lemmatisation** : association d'un lemme à chaque mot
- Lemme : unité autonome constituant le lexique d'une langue (= une entrée du dictionnaire)
  - Chiennes → chien
  - Mangerait → manger
  - Président → présider OU président ?

- **Étiquetage morpho-syntaxique** (POS tagging) : association d'informations grammaticales aux mots (nature, genre, nombre…)

- Quelques outils :
  - TreeTagger
  - LIA Tagg
  - Stanford Tagger

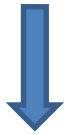
- Efficacité : ~97% (corpus généralistes)

TreeTagger : 33 tags en français

Tag	Signification	Tag	Signification
ABR	abreviation	PRP:det	preposition plus article (au,du,aux,des)
ADJ	adjective	PUN	punctuation
ADV	adverb	PUN:cit	punctuation citation
DET:ART	article	SENT	sentence tag
DET:POS	possessive pronoun (ma, ta, ...)	SYM	symbol
INT	interjection	VER:cond	verb conditional
KON	conjunction	VER:futu	verb futur
NAM	proper name	VER:impe	verb imperative
NOM	noun	VER:impf	verb imperfect
NUM	numeral	VER:infi	verb infinitive
PRO	pronoun	VER:pper	verb past participle
PRO:DEM	demonstrative pronoun	VER:ppre	verb present participle
PRO:IND	indefinite pronoun	VER:pres	verb present
PRO:PER	personal pronoun	VER:simp	verb simple past
PRO:POS	possessive pronoun (mien, tien, ...)	VER:subi	verb subjunctive imperfect
PRO:REL	relative pronoun	VER:subp	verb subjunctive present
PRP	preposition		

TreeTagger : exemple

« TreeTagger permet d'annoter plusieurs langues. »



Mot	POS	Lemme
TreeTagger	NAM	<unknown>
permet	VER:pres	permettre
d'	PRP	de
annoter	VER:infi	annoter
plusieurs	PRO:IND	plusieurs
langues	NOM	langue
.	SENT	.

Plan du cours

1. Introduction

2. Analyse lexicale

3. **Analyse syntaxique**
   - Grammaire hors contexte
   - Analyses montante et descendante
   - Treebanks
   - Robustesse et analyse syntaxique de surface

4. Analyse sémantique

Analyse syntaxique : principes

- Toute suite de mots n'est pas acceptable :
  - Verte mange Jean pomme la

- **Syntaxe** : étude des contraintes définissant les successions licites de formes (i.e. phrase grammaticalement correctes)
- Constituants : groupes de mots se comportant comme une seule unité
  - Groupes nominaux : Ils, Jean, ceci… la maison, le petit chat gris…
- Relations grammaticales :
  - Jean [SUJET] mange la pomme verte [OBJET]

- Nécessité d'une grammaire définie de manière formelle

Grammaire hors contexte

- Context Free Grammar (CFG) :
  - ensemble de **règles** exprimant la manière dont les symboles d'un langage peuvent être groupés et agencés ensemble, sous la forme X → Y
  - lexique de mots et symboles :
    - **Terminaux** : mots du langage
    - **Non-terminaux** : regroupements ou généralisations

Note : X → Y, X = non-terminal, Y = liste ordonnée d'un ou plusieurs terminaux et non-terminaux

- Peut-être utilisée pour :
  - Associer une structure à une phrase donnée
  - Générer des phrases

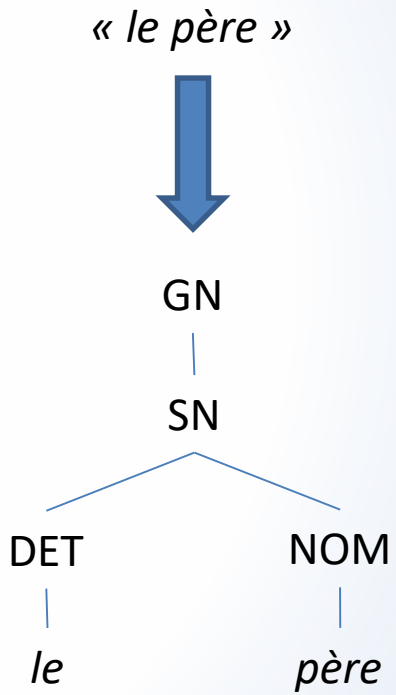
CFG : exemple

Exemple de grammaire :

1. Règles :
   - SN → DET NOM
   - SN → NAM (nom propre)
   - GN → SN
     | GN PUN SN
     | GN KON SN

2. Terminaux :
{Jean, Pierre, le, la, son, sa, père, mère, et}

GN = symbole de départ



Arbre d'analyse

CFG : exemple (2)

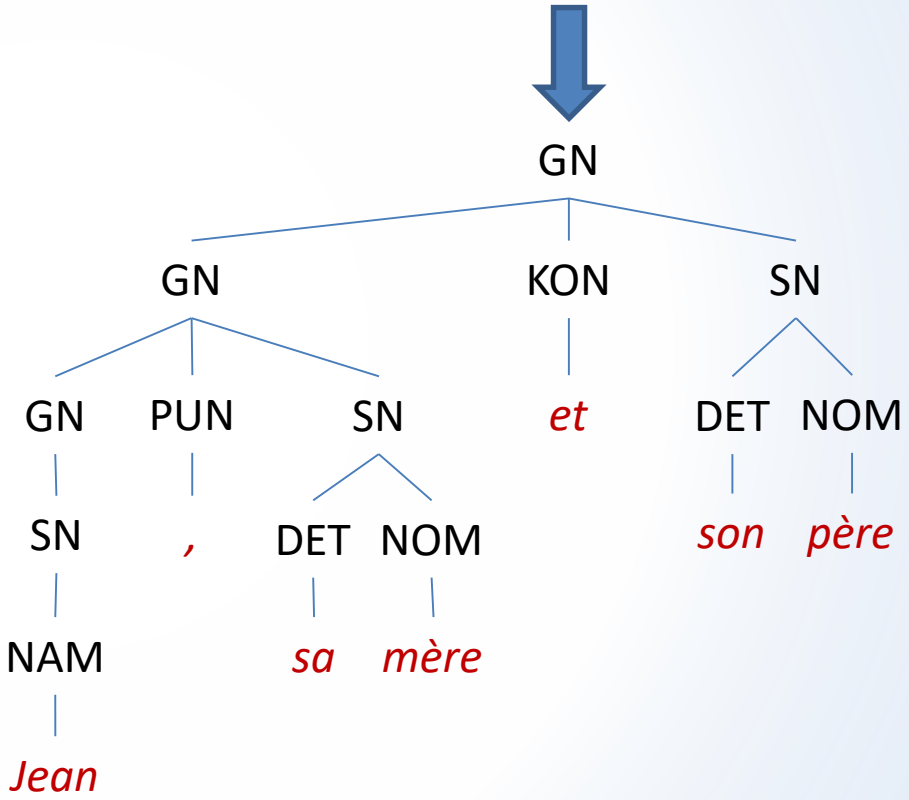
« Jean, son père et sa mère »

Exemple de grammaire :

1. Règles :
   - SN → DET NOM
   - SN → NAM (nom propre)
   - GN → SN
     | GN PUN SN
     | GN KON SN

2. Terminaux :
{Jean, Pierre, le, la, son, sa, père, mère, et}

GN = symbole de départ



OU

[GN [GN [GN [SN [NAM Jean]]] [PUN ,] [SN [DET sa] [NOM mère]]] [KON et] [SN [DET son][NOM père]]]

CFG : exemple (3)

Exemple de grammaire :

1. Règles :
   - SN → DET NOM
   - SN → NAM (nom propre)
   - GN → SN
     - | GN PUN SN
     - | GN KON SN

2. Terminaux :
{Jean, Pierre, le, la, son, sa, père, mère, et}

GN = symbole de départ

Phrases pouvant être générées :

Jean
Le père
Sa mère
Jean, Pierre et le père
…

Mais aussi :
La père
Son mère
Jean, Pierre, le père, sa mère, Pierre
Jean, Jean, Jean, Jean, Jean et Jean
…

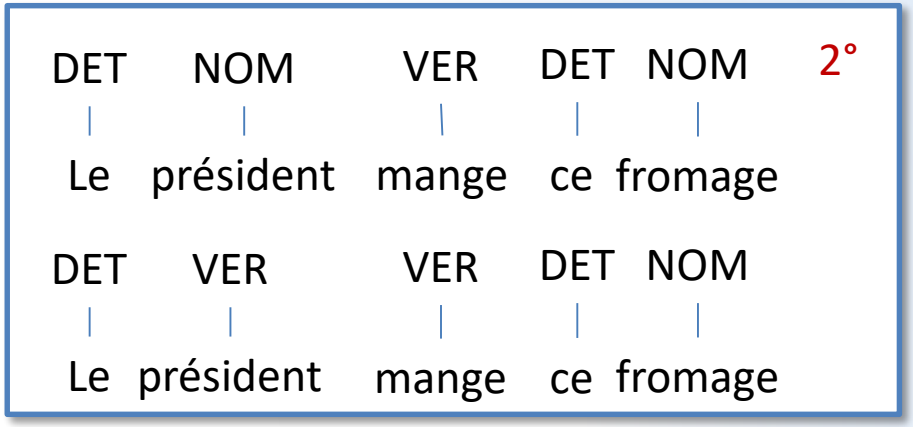
Analyse montante vs. descendante

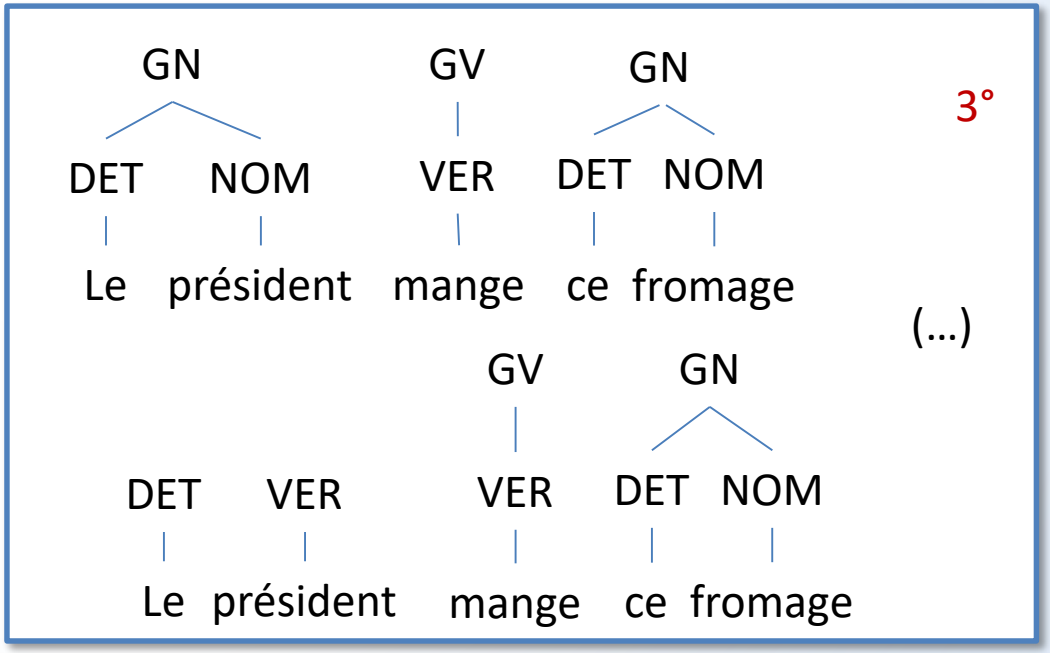
- Analyse **montante** (bottom-up) : départ des mots composant la phrase et tentative de remontée à la racine

- Grammaire :
  - S → GN GV | GV
  - GN → PRO | DET NOM
  - GV → VER | VER GN
  - PRO → il | elle
  - NOM → président | fromage
  - DET → le | ce
  - VER → manger | présider

- Phrase :

« Le président mange ce fromage »





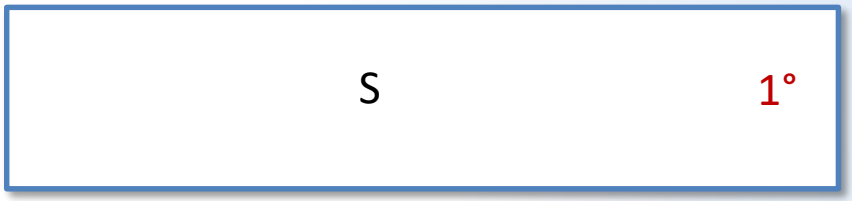
Analyse montante vs. descendante

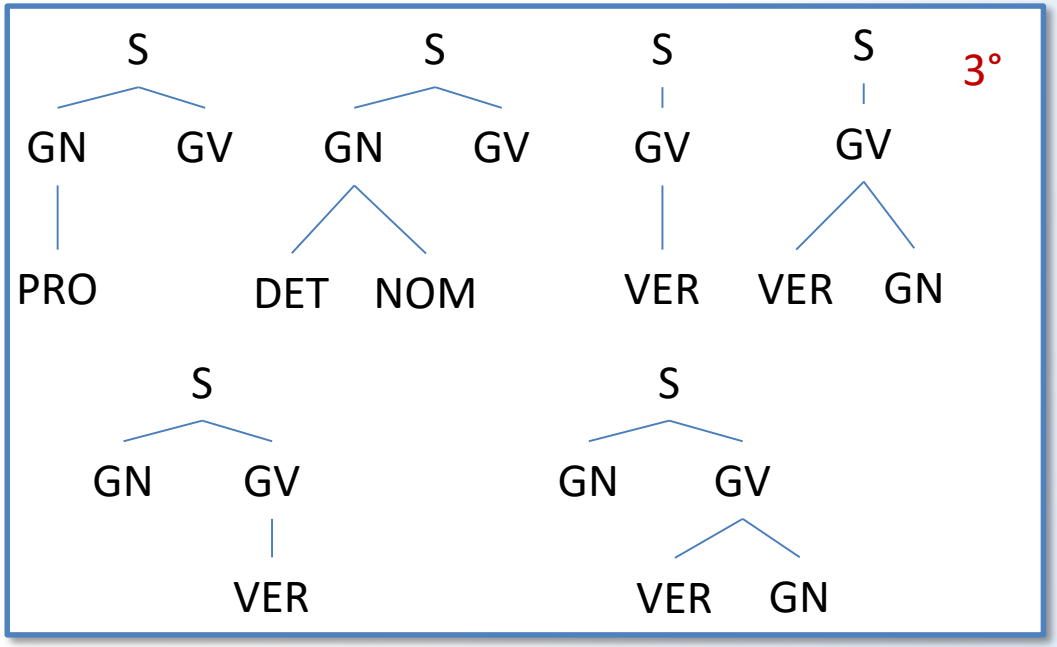
- Analyse **descendante** (top-down) : départ du nœud racine puis expansion jusqu'à atteindre la phrase

- Grammaire :
  - S → GN GV | GV
  - GN → PRO | DET NOM
  - GV → VER | VER GN
  - PRO → il | elle
  - NOM → président | fromage
  - DET → le | ce
  - VER → manger | présider

- Phrase :

« Le président mange ce fromage »





Treebanks

- **Treebank :** corpus de textes annotés syntaxiquement à l'aide de grammaires
- Utile pour :
  - évaluer des étiqueteurs morpho-syntaxiques, grammaires et analyseurs syntaxiques
  - extraire des règles de production
- En anglais : **Penn Treebank**
  - 4.5M de mots
  - articles de presse, livres scientifiques et de fiction
  - 4500 règles utilisées
  - annotations à la main
- En français : **French Treebank**
  - 24000 phrases (780 000 mots)
  - articles du Monde entre 1989 et 1995

Exemple de phrase issue du French Treebank :

```
<SENT>
  <PP fct="MOD">Au <NP>début</NP></PP>,
  <VN fct="SUJ">on ramassait</VN>
  <VPinf fct="OBJ">
    <PP fct="DE-OBJ">de <NP>quoi</NP></PP>
    <VN>remplir</VN>
    <NP fct="OBJ">quinze sacs_poubelle</NP>
  </VPinf>,
  <Sint>
    <VN>indique</VN>
    <NP fct="SUJ">Roger,
      <NP>ouvrier <PP>? <NP>la régie</NP></PP></NP>
    </NP>
  </Sint>.
</SENT>
```

Analyse syntaxique : robustesse

- Problèmes :
  - Le nombre d'arbres de dérivation peut exploser rapidement (et le temps de calcul avec) : parfois des milliers pour une phrase complexe (dont des redondantes)
  - Absence d'une structure = 0 arbre
  - Pas efficace sur langue en situations réelles d'interaction (phrases non grammaticales)

- **Robustesse** : capacité à produire des analyses utiles (i.e. au moins partiellement correctes) pour des textes réels

Analyse syntaxique de surface (shallow parsing)

- Idées-clé :
  - Limitation de la profondeur d'analyse (analyses partielles)
  - Production d'une **sortie unique** pour chaque entrée
  - Reconnaissance de syntagmes noyaux (chunks) linguistiquement motivés (SN, SV…)

- Exemple : cascade de transducteurs
  - Analyse par îlots de confiance
  - Définition de niveaux (ensemble de règles) :
    - L'entrée d'un niveau est la sortie du précédent
    - Règles les plus sûres à un niveau bas

- Efficacité : ~90% analyse de surface (corpus généralistes)

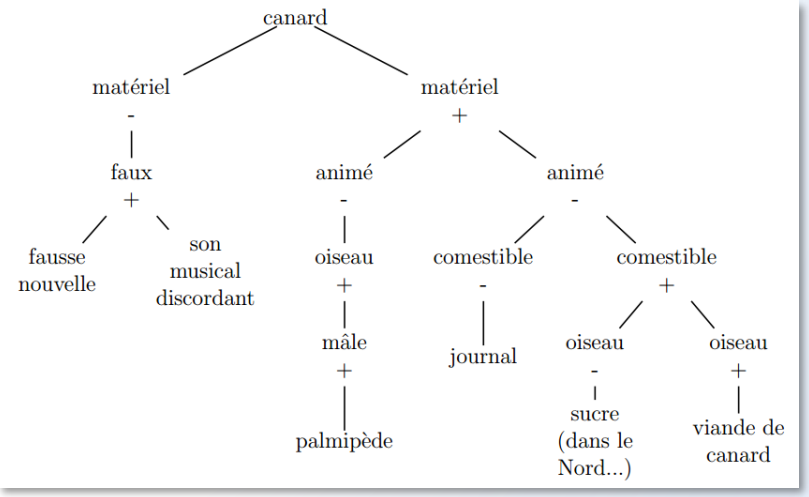
Plan du cours

1. Introduction
2. Analyse lexicale
3. Analyse syntaxique
4. **Analyse sémantique**
   - Sémantique lexicale
     - Sèmes
     - WordNet
   - Sémantique propositionnelle
     - FrameNet
     - Logique du Premier Ordre
     - Graphes conceptuels

Analyse sémique

- **Sème** : unité minimale de signification, dont les valeurs possibles sont +, -, Ø
- Jument, poulain et pouliche ont des sèmes communs : ils appartiennent au même **champ sémantique**.
- Peut permettre de distinguer différents sens d'un même mot
- Limites :
  - Consensus difficile
  - Pas d'ensemble de tous les sèmes possibles existant dans la langue générale
  - Adapté aux mots lexicaux (!= par, qui, dont)

	cheval	mâle	adulte
Jument	+	-	+
Poulain	+	+	-
Pouliche	+	-	-



Sémantique lexicale : principes

- Comment attacher un sens aux mots ?
- Dans le dictionnaire, les définitions sont circulaires :
  - **Rouge** : de la couleur du **sang** ou du rubis
  - **Sang** : liquide **rouge** qui circule dans le cœur, les artères et les veines des animaux

- Besoin d'encoder les relations de sens entre mots dans des bases de données :
  - **Synonymie** : sens similaire à celui d'origine
    canapé / sofa, voiture / automobile
  - **Antonymie** : sens opposé à celui d'origine
    court / long, haut / bas, clair / sombre
  - **Hyponymie** : sens plus spécifique par rapport à celui d'origine : voiture pour véhicule, chien pour animal
  - **Hyperonymie** : sens plus général par rapport à celui d'origine : meuble pour chaise, fruit pour pomme

Sémantique lexicale : WordNet

- WordNet [Fellbaum, 1998] : base de données lexicales en anglais
  - 3 bases de données :
    - Noms (> 117k), 1.23 sens /nom en moyenne
    - Verbes (> 11k), 2.16 sens / verbe en moyenne
    - Adjectifs (> 22k) & adverbes (> 4k)
  - Associations **synsets** et gloses
  - Encode relations entre synsets
  - Accessible depuis le web ou en local
- Il existe des équivalents (moins riches) pour d'autres langues : EuroWordNet

BASS (n):
1. bass[1] – the lowest part of the musical range
2. bass[2], basso[1] – an adult male singer with the lowest voice
3. seabass[1], bass[3] – the lean flesh of a saltwater fish of the family Serranidae
4. bass[5], bassvoice[1], basso[2] – the lowest adult male singing voice

(+ 4 others)

BASS (v):
1. bass[1], deep[6] – (having or denoting a low vocal or instrumental range)

Word Sense Disambiguation (WSD) :
if tenor is close → bass[2]
if fish is close → bass[3]

Sémantique propositionnelle : FrameNet

- FrameNet [Baker et al., 1998; Ruppenhofer et al., 2006] :
  - **Frame** = une structure scriptée avec des rôles sémantiques associés
  - Rôles sémantiques : peuvent être principaux ou optionnels

- Exemple : changement de valeur sur une échelle

  « Cette frame représente les mots indiquant un changement de la position d'un objet sur une échelle (l'Attribut) depuis une valeur (Valeur_Initiale) à une autre (Valeur_finale). »

  [$_{\text{ATTRIBUTE}}$ Le prix] [$_{\text{ITEM}}$ des bananes] a augmenté de [$_{\text{DIFFERENCE}}$ 2%].

  [$_{\text{ITEM}}$ Son cours] est descendu à [$_{\text{FINAL\_VALUE}}$ $42].

  [$_{\text{ITEM}}$ La température] a [$_{\text{SPEED}}$ rapidement] chuté de [$_{\text{INITIAL\_VALUE}}$ 23°C] à [$_{\text{FINAL\_VALUE}}$ 12 °C] en [$_{\text{DURATION}}$ 24 heures].

Sémantique propositionnelle : mise en pratique

- Comment modéliser « cliquer » avec une approche de type frame ?

1. Collecter un corpus en situation
2. Rechercher les occurrences de cliquer
3. En déduire les rôles sémantiques associés

Schéma DAFT pour cliquer

```
<schema name="Click">
 <class>A</class>
 <type>Act</type>
 <info>A person click on an object
 in a certain manner</info>
 <fields>
  <field type="person" attribute="clicker" />
  <field type="object" attribute="clicked" />
  <field type="manner" attribute="manner" />
 </fields>
</schema>
```

[Bouchet & Sansonnet, 2010]

Logique du premier ordre (LPO)

- Objectif : permettre de **représenter les connaissances** du monde et de raisonner sur des propositions

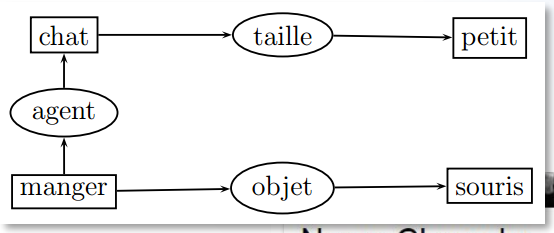
- Exemples :
  - Minou est un chat :
    chat(Minou)
  - Minou ne dort pas :
    $$\neg dormir(Minou)$$
  - Minou regarde Jean :
    $$regarde(Minou, Jean)$$
  - « un chat dort » :
    $$\exists x\ (chat(x) \wedge dort(x))$$
  - « tous les chats sont des félins » :
    $$\forall x\ (chat(x) \rightarrow felix(x))$$

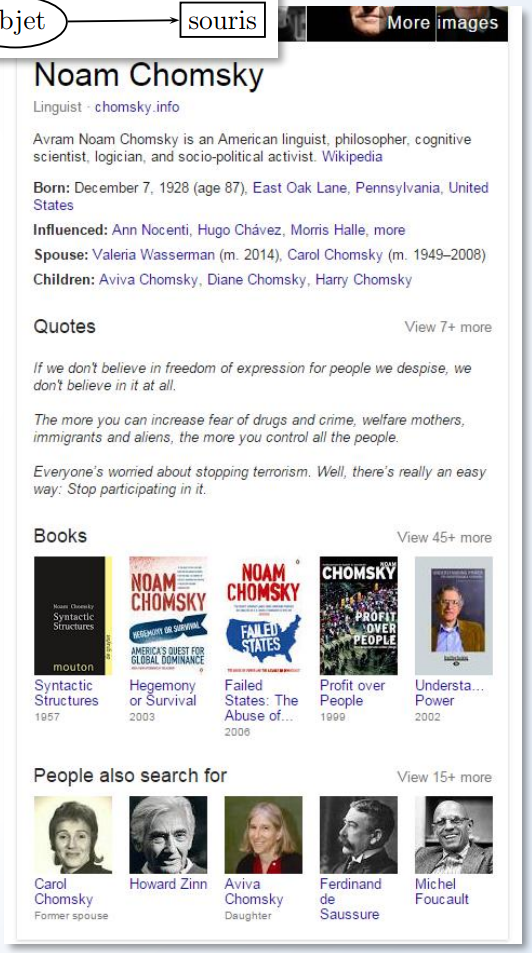
Graphes conceptuels

- Alternative à la LPO, sans formules logiques, mais permettant néanmoins le raisonnement

- Exemple : « le petit chat mange une souris »

- Exemple 2 : Google Knowledge Graph





Limites des approches sémantiques propositionnelles (présentées)

- Sens commun de « et » et « ou » :
  - J'aime lire **ou** aller au cinéma
- Il faut représenter tout ce qui existe dans le monde
- Problème des entités nommées
- Figures de style difficile à modéliser (pas d'humour ou d'ironie)

Conclusion

- L'analyse d'un texte requiert une série d'analyses complexes, et les erreurs à chaque étape affectent l'étape suivante
- Possibilité de se concentrer sur un niveau inférieur et utiliser niveaux supérieurs pour amélioration (e.g. text mining sur des mots)
- Nombreux outils existent (lemmatisation, étiquetage morpho-syntaxique, analyse syntaxique…) & bibliothèques associées (NLTK pour Python, OpenNLP pour Java…)
- Pour obtenir de bonnes performances :
  - Mieux vaut se concentrer sur un **sous-domaine** de la langue
    → Pas de passage à l'échelle des analyses profondes
  - Collecter, si possible, un **corpus** en situation réelle