Quelques bases de Traitement Automatique **des** Langues (TAL)

Karën Fort

karen.fort@univ-lorraine.fr / https://members.loria.fr/KFort/

Institut des sciences du Digital Management & Cognition
COMPOSANTE DE L'UNIVERSITÉ DE LORRAINE

UNIVERSITÉ DE LORRAINE

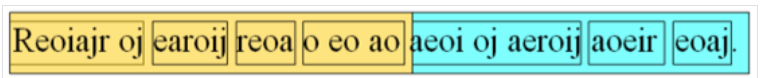
Quelques sources d'inspiration

- Cours d'Alain Couillault, Gestion sémantique des contenus, Master ICONE, La Rochelle.
- http://faculty.washington.edu/ebender/2012_472/0404.pdf
- http://www.cis.upenn.edu/~cis639/docs/fsexamples.html

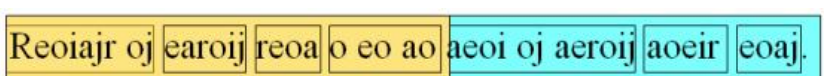
De quoi ça parle ?

Reoiajr oj earoij reoa o eo ao aeoi oj aeroij aoeir eoaj.

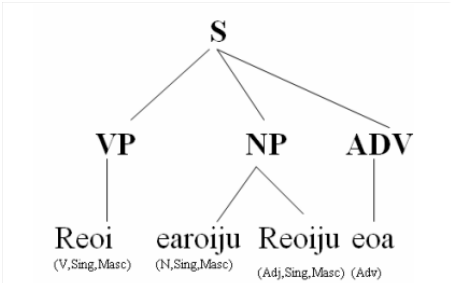
De quoi ça parle ?

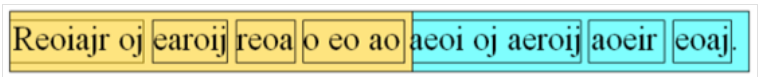


De quoi ça parle ?

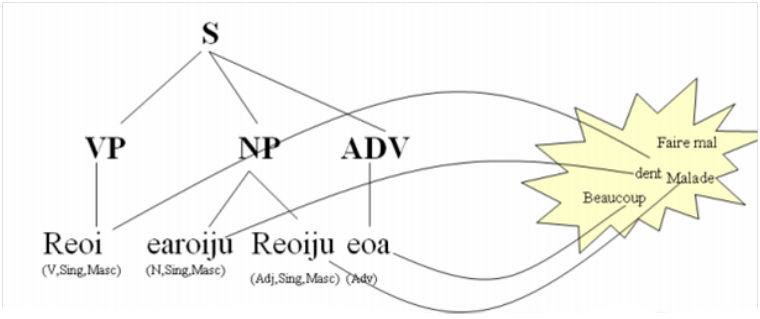


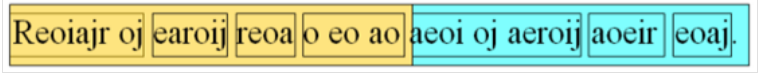
De quoi ça parle ?





De quoi ça parle ?





Pour quoi faire ?

- traduction automatique
- recherche d'informations
- indexation
- traitement de courriels
- veille
- agents conversationnels
- ...

Associations savantes

Acteurs majeurs

(quelques) Entreprises de TAL en France

Marché (tel que vu il y a quelques années)

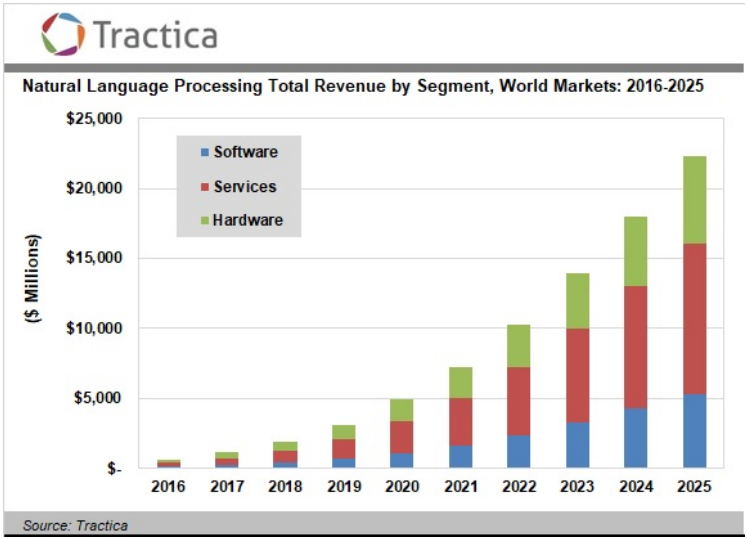
"The NLP market size, which is about $7.5B today, is estimated to grow to $16B by 2021."

https://www.techemergence.com/natural-language-processing-business-applications/

"The next few years should see AI technology increase even more, with the global AI market expected to push $60 billion by 2025"

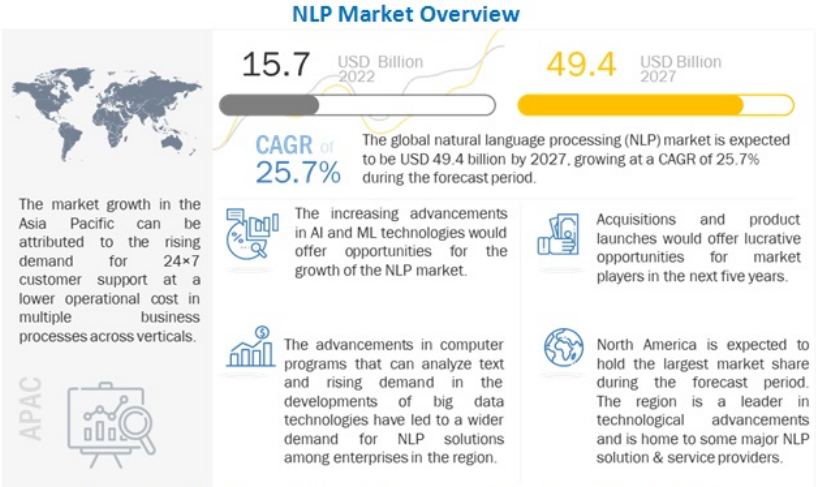
https://www.forbes.com/sites/forbestechcouncil/2018/07/02/what-is-natural-language-processing-and-what-is-it-used-for/#7ef6bf675d71

Prévisions de chiffre d'affaire (il y a quelques années)



https://www.tractica.com/research/natural-language-processing/

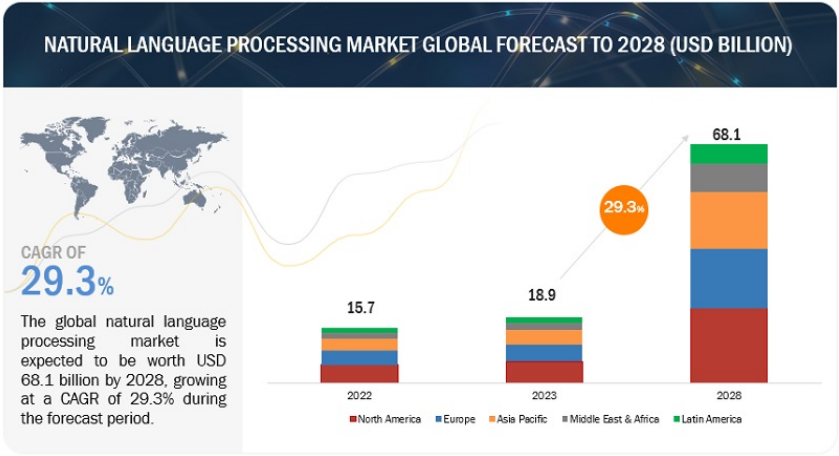
Prévisions (il y a 2 ans)



Source: Secondary Research, Expert Interviews, and MarketsandMarkets Analysis

https://www.marketsandmarkets.com/Market-Reports/natural-language-processing-nlp-825.html

Prévisions aujourd'hui (2024)



https://www.marketsandmarkets.com/Market-Reports/natural-language-processing-nlp-825.html

TAL et linguistique computationnelle

Linguistique computationnelle :

- modéliser une activité langagière pour comprendre comment fonctionne la langue

vs

Traitement automatique des langues :

- reproduire l'activité langagière humaine pour optimiser les performances des systèmes

Le TAL : pas si simple...

- *La mousse aux fraises est sur la table de l'avocat.*

Le TAL : pas si simple...

- *La mousse aux fraises est sur la table de l'avocat.*
- *L'omelette au lard est parti sans payer !*

Le TAL : pas si simple...

- *La mousse aux fraises est sur la table de l'avocat.*
- *L'omelette au lard est parti sans payer !*
- *Le bus a renversé un passant...*

Le TAL : pas si simple...

- *La mousse aux fraises est sur la table de l'avocat.*
- *L'omelette au lard est parti sans payer !*
- *Le bus a renversé un passant...*
  - *... je l'ai entendu freiner.*

Le TAL : pas si simple...

- *La mousse aux fraises est sur la table de l'avocat.*
- *L'omelette au lard est parti sans payer !*
- *Le bus a renversé un passant...*
  - *... je l'ai entendu freiner.*
  - *... je l'ai entendu crier.*

Le TAL : pas si simple...

- *La mousse aux fraises est sur la table de l'avocat.*
- *L'omelette au lard est parti sans payer !*
- *Le bus a renversé un passant...*
  - *... je l'ai entendu freiner.*
  - *... je l'ai entendu crier.*
- *Le professeur a envoyé l'élève chez le proviseur...*

Le TAL : pas si simple...

- *La mousse aux fraises est sur la table de l'avocat.*
- *L'omelette au lard est parti sans payer !*
- *Le bus a renversé un passant...*
  - *... je l'ai entendu freiner.*
  - *... je l'ai entendu crier.*
- *Le professeur a envoyé l'élève chez le proviseur...*
  - *... il faisait trop de bruit.*

Le TAL : pas si simple...

- *La mousse aux fraises est sur la table de l'avocat.*
- *L'omelette au lard est parti sans payer !*
- *Le bus a renversé un passant...*
  - *... je l'ai entendu freiner.*
  - *... je l'ai entendu crier.*
- *Le professeur a envoyé l'élève chez le proviseur...*
  - *... il faisait trop de bruit.*
  - *... il était excédé.*

Le TAL : pas si simple...

- *La mousse aux fraises est sur la table de l'avocat.*
- *L'omelette au lard est parti sans payer !*
- *Le bus a renversé un passant...*
  - *... je l'ai entendu freiner.*
  - *... je l'ai entendu crier.*
- *Le professeur a envoyé l'élève chez le proviseur...*
  - *... il faisait trop de bruit.*
  - *... il était excédé.*
  - *... il l'avait convoqué.*

Le TAL : pas si simple...

- *La mousse aux fraises est sur la table de l'avocat.*
- *L'omelette au lard est parti sans payer !*
- *Le bus a renversé un passant...*
  - *... je l'ai entendu freiner.*
  - *... je l'ai entendu crier.*
- *Le professeur a envoyé l'élève chez le proviseur...*
  - *... il faisait trop de bruit.*
  - *... il était excédé.*
  - *... il l'avait convoqué.*

→ ambiguïtés pour les systèmes et/ou pour les humains

Le découpage en « mots » ou *tokenization*

- *l'arbre*

Le découpage en « mots » ou *tokenization*

- *l'arbre*
- *aujourd'hui*

L'analyse morphologique

la *porte*

- *porte* +Nf + Sg

L'analyse morphologique

la *porte*

- *porte* +Nf + Sg
- *porte* +VT + 1/3P + Sg

L'analyse syntaxique

Jean regarde un homme sur la colline avec un télescope.

- **Qui** est sur la colline ?
- **Qui** a un télescope ?

L'analyse sémantique

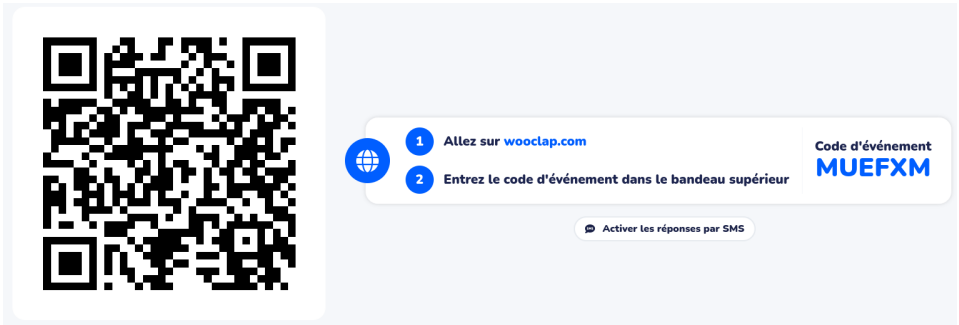
Tous les hommes aiment une femme.

→ **Chaque** homme aime une femme ou **tous** les hommes aiment la même femme ?

Des techniques variées

- systèmes à base de règles
  - définies par l'humain (linguistes/info-linguistes)
  - entrées manuellement
- systèmes basés sur les données
  - apprentissage supervisé ou non supervisé
  - à partir d'exemples (rédigés et/ou annotés par des humains)
  - algorithmes (pensés par des humains)

Quelles techniques dans l'industrie ?



Exercice : identifier la langue d'un texte

Groupes de 3

Exercice

Trouver au moins **2** algorithmes permettant d'identifier la langue d'un texte

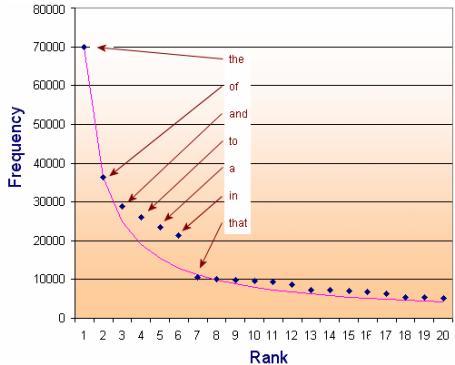
Conseil : connaître les langues et leurs caractéristiques

Loi de Zipf :

- étant donné un corpus d'énoncés en langue naturelle, la fréquence d'un mot est **inversement proportionnelle** à son rang dans la table de fréquence.
- ex. : "the" représente près de 7 % du *Brown Corpus* alors que près de la moitié du vocabulaire total du corpus sont des hapax.
- Seuls 135 éléments de vocabulaire sont nécessaires pour couvrir la moitié du *Brown Corpus*

Rang	Mot	Fréquence
1	*the*	69 970
2	*of*	36 410
3	*and*	28 854
20	*I*	5 180

Loi de Zipf sur le *Brown corpus*



Identifier la langue : solution 1

Méthode des *shortwords* :

- BD de « mots outils » (mots grammaticaux, « petits » mots)
- Compter les occurrences de ces mots outils dans le texte
- Comparer avec les BD

Identifier la langue : solution 2

Méthode des *trigrammes* :

- Rechercher la probabilité qu'un caractère C_i apparaisse après les deux précédents dans la langue l :

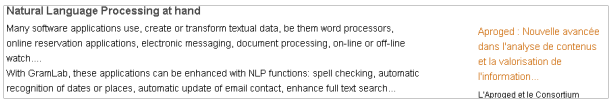
$$P(C_i|C_{i-1} : C_{i-2}, l)$$

- Calculer la probabilité résultante pour chaque langue, pour l'ensemble du texte :

$$\prod_{i=1}^{i=n} P(C_i|C_{i-1} : C_{i-2}, l)$$

Limitations

- Longueur du texte (5 mots mini.)
- textes multilingues :



- et aussi :
  - Barack Obama $\rightarrow$ français
  - Barack Obama and Nicolas Sarkozy $\rightarrow$ anglais
  - camping caravaning, trekking

Identifier la langue : solutions 3 et 4

- Essayer d'identifier l'encodage
- Regarder les méta-données

Google language identifier

https://translate.google.com/?sl=auto

Parenthèse : à propos des mots outils (*stopwords*)

À quoi servent-ils ? Qu'en faire ?

Python3

```python
import nltk
from nltk.corpus import stopwords

nltk.download('stopwords')
print(stopwords.words('english'))
```

https://www.geeksforgeeks.org/removing-stop-words-nltk-python/

Ce que vous enlevez quand vous enlevez les mots outils

```
108  some
109  such
110  no
111  nor
112  not
113  only
114  own
115  same
116  so
117  than
118  too
```

https://gist.github.com/sebleier/554280#file-nltk-s-list-of-english-stopwords

Segmentations : exercice

Marchepied à plate forme sécurisée 2.41m hauteur travail max. 3.16m 5 marches et garde corps fixe PROFORT TUBESCA

Exercice

Segmenter (manuellement) la phrase en mots

Segmentations : approche simpl(ist)e

1. Segmentation en phrases (texte brut) :
   - Se termine par une ponctuation de fin de phrase (. ; ? !)
   - La phrase suivante commence par une espace ou un retour à la ligne suivi d'une majuscule
   - Les ... forment une ponctuation unique
2. Segmentation en mots (tokenization) :
   - Séparateurs de mots : espace, CR, tab, apostrophes (avec le mot), ponctuations (considérés comme des mots), les . ne sont pas des séparateurs
   - Les tirets sont des séparateurs s'ils sont suivis d'un pronom (vient-il), et la séquence –t doit être effacée dans certains cas (*envoie-t-elle*, césure de fin de phrase)
   - Appliquer des transformations locales : Au, du, cet, qu', l', m'. . ...

Segmentation en phrases : plus loin

- en allemand ?
- en thaï ?

Tokenization : plus loin

Je veux bien que tu viennes

→ Conserver l'ambiguïté

Tokenization : encore plus loin

- en allemand ?
- en chinois ?

Segmentations : les questions

Définition théorique du segment :

- Qu'est-ce qu'un mot ?
  - Plate forme, marche-pied marche pied marchepied
  - Les marchands du temple : du → de le ?
- Qu'est-ce qu'une phrase ?
  - Ah ?
  - - ?

Ambiguïté des séparateurs :

- Microsoft.com, 23.5, . . ., C.G.T.
- Aujourd'hui, 9'8, jusqu'à
- C&A, R&D
- Sépara-
- teur

Parenthèse sur la tokenization dans les grands modèles de langues

Ici sur l'interface de test du tokenizer d'OpenAI (H. de Mazancourt)

https://www.linkedin.com/in/mazancourt/

Morphologie : exercice

Orange, au cours du matin, a permis la prise de bénéfices. Nous avions fait des paris audacieux sur cette valeur qui est montée dès l'ouverture du CO

Article, Nom, Adverbe, Verbe, Adjectif, Pronom, Nom propre, Ppassé, Préposition, Interjection

Exercice

Utiliser les étiquettes proposées pour réaliser une analyse morpho-syntaxique du texte

Morphologie : étiquettes

- Orange [Nom propre, Adjectif, Nom]
- au [Préposition, Article]
- cours [Nom]
- au cours du [Préposition]
- du [Ppassé, Nom]
- matin [interjection, Adverbe, Nom]
- a [Préposition, Nom]
- permis [Ppassé, Nom, Adjectif]
- la [Article, Pronom, Nom]
- prise [Nom, Ppassé]
- de [Préposition]
- bénéfices [Nom]

Morphologie : lemmatisation

- Orange → orange — Orange
- au → à + le
- cours → cours
- du → de + le
- matin → matin
- a → avoir
- permis → permis — permettre
- la → le
- prise → prise — prendre
- de → de
- bénéfices → bénéfice

Morphologie : solutions en extension

```
<lexicalEntry id="championne_1">
  <feminineVariantOf target="champion_1">champion</feminineVariantOf>
     <formSet>
        <lemmatizedForm>
           <orthography>championne</orthography>
          <grammaticalCategory>commonNoun</grammaticalCategory>
          <grammaticalGender>feminine</grammaticalGender>
        </lemmatizedForm>
        <inflectedForm>
           <orthography>championne</orthography>
           <grammaticalNumber>singular</grammaticalNumber>
        </inflectedForm>
        <inflectedForm>
           <orthography>championnes</orthography>
           <grammaticalNumber>plural</grammaticalNumber>
        </inflectedForm>
    </formSet>
  <originatingEntry target="TLF">CHAMPION, ONNE, subst.</originatingEntry>
</lexicalEntry>
```

Stockage ?

Morphalou 2 = **160 Mo**

Temps d'accès ?

Morphologie : la composition

https://www.youtube.com/watch?v=gG62zay3kck

Morphologie : l'agglutination

Exemples en finnois :

- *Järjestelmällinen* : organisé

Morphologie : l'agglutination

Exemples en finnois :

- *Järjestelmällinen* : organisé
- *järjestelmällistetty* : conçu pour être organisé

Morphologie : l'agglutination

Exemples en finnois :

- *Järjestelmällinen* : organisé
- *järjestelmällistetty* : conçu pour être organisé
- *epäjärjestelmällistetty* : désorganisé ou pas conçu pour être organisé

Morphologie : l'agglutination

Exemples en finnois :

- *Järjestelmällinen* : organisé
- *järjestelmällistetty* : conçu pour être organisé
- *epäjärjestelmällistetty* : désorganisé ou pas conçu pour être organisé
- *Epäjärjestelmällistämättömyy* : négation et substantivation

Morphologie : l'agglutination

Exemples en finnois :

- *Järjestelmällinen* : organisé
- *järjestelmällistetty* : conçu pour être organisé
- *epäjärjestelmällistetty* : désorganisé ou pas conçu pour être organisé
- *Epäjärjestelmällistämättömyy* : négation et substantivation
- *Epäjärjestelmällistyttämättömyydellänsä* : l'objet dénoté par le substantif appartient à quelqu'un (génitif)

Morphologie : l'agglutination

Exemples en finnois :

- *Järjestelmällinen* : organisé
- *järjestelmällistetty* : conçu pour être organisé
- *epäjärjestelmällistetty* : désorganisé ou pas conçu pour être organisé
- *Epäjärjestelmällistämättömyy* : négation et substantivation
- *Epäjärjestelmällistyttämättömyydellänsä* : l'objet dénoté par le substantif appartient à quelqu'un (génitif)
- *Epäjärjestelmällistyttämättömyydellänsäkään* : nouvelle négation

Morphologie : l'agglutination

Exemples en finnois :

- *Järjestelmällinen* : organisé
- *järjestelmällistetty* : conçu pour être organisé
- *epäjärjestelmällistetty* : désorganisé ou pas conçu pour être organisé
- *Epäjärjestelmällistämättömyy* : négation et substantivation
- *Epäjärjestelmällistyttämättömyydellänsä* : l'objet dénoté par le substantif appartient à quelqu'un (génitif)
- *Epäjärjestelmällistyttämättömyydellänsäkään* : nouvelle négation
- *epäjärjestelmällistyttämättömyydellänsäkäänkö* : ajout d'un suffixe dénotant l'interrogation

Morphologie : l'agglutination

Exemples en finnois :

- *Järjestelmällinen* : organisé
- *järjestelmällistetty* : conçu pour être organisé
- *epäjärjestelmällistetty* : désorganisé ou pas conçu pour être organisé
- *Epäjärjestelmällistämättömyy* : négation et substantivation
- *Epäjärjestelmällistyttämättömyydellänsä* : l'objet dénoté par le substantif appartient à quelqu'un (génitif)
- *Epäjärjestelmällistyttämättömyydellänsäkään* : nouvelle négation
- *epäjärjestelmällistyttämättömyydellänsäkäänkö* : ajout d'un suffixe dénotant l'interrogation
- *epäjärjestelmällistyttämättömyydellänsäkäänköhän* : ajout d'une emphase (c'est moi qui ...)

Morphologie : l'agglutination

Exemples en finnois :

- *Järjestelmällinen* : organisé
- *järjestelmällistetty* : conçu pour être organisé
- *epäjärjestelmällistetty* : désorganisé ou pas conçu pour être organisé
- *Epäjärjestelmällistämättömyy* : négation et substantivation
- *Epäjärjestelmällistyttämättömyydellänsä* : l'objet dénoté par le substantif appartient à quelqu'un (génitif)
- *Epäjärjestelmällistyttämättömyydellänsäkään* : nouvelle négation
- *epäjärjestelmällistyttämättömyydellänsäkäänkö* : ajout d'un suffixe dénotant l'interrogation
- *epäjärjestelmällistyttämättömyydellänsäkäänköhän* : ajout d'une emphase (c'est moi qui ...)
- *epäjärjestelmällistyttämättömyydelläänsäkäänköhän* :

Morphologie : l'agglutination

Exemples en finnois :

- *Järjestelmällinen* : organisé
- *järjestelmällistetty* : conçu pour être organisé
- *epäjärjestelmällistetty* : désorganisé ou pas conçu pour être organisé
- *Epäjärjestelmällistämättömyy* : négation et substantivation
- *Epäjärjestelmällistyttämättömyydellänsä* : l'objet dénoté par le substantif appartient à quelqu'un (génitif)
- *Epäjärjestelmällistyttämättömyydellänsäkään* : nouvelle négation
- *epäjärjestelmällistyttämättömyydellänsäkäänkö* : ajout d'un suffixe dénotant l'interrogation
- *epäjärjestelmällistyttämättömyydellänsäkäänköhän* : ajout d'une emphase (c'est moi qui ...)
- *epäjärjestelmällistyttämättömyydelläänsäkäänköhän* :

> *The reverse of the reverse of something abstract that is made to be unorganised, which is owned by someone, and is one of the two or more (possibly similar) attributes that have a negative atmosphere or lack of something, and we doubt if it is it at the same time that we ensure that it truly is.*

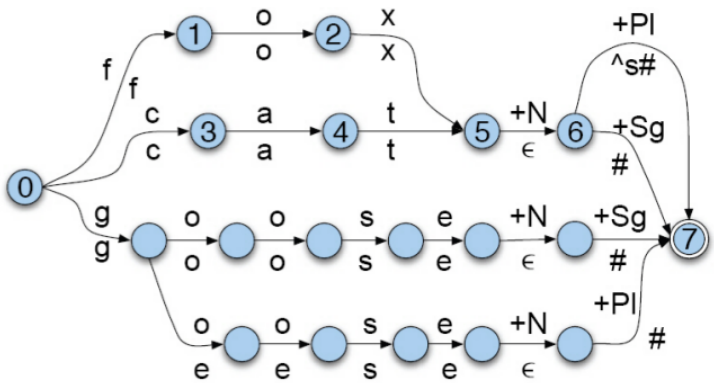
Morphologie : solutions en intension

Quelle solution pour :

- optimiser le stockage
- permettre un accès rapide
- dans les deux sens :
  - leave+NNS $\rightarrow$ leaves
  - leaves $\rightarrow$ leave+NNS

Structure de données **et** accès ?

Morphologie : solutions en intension



Morphologie : transducteurs à états finis (FST)

XFST (Xerox FST) :

```
[ [l e a v e %+VBZ .x. l e a v e s] |
[l e a v e %+VB .x. l e a v e] |
[l e a v e %+VBG .x. l e a v i n g] |
[l e a v e %+VBD .x. l e f t] |
[l e a v e %+NN .x. l e a v e] |
[l e a v e %+NNS .x. l e a v e s] |
[l e a f %+NNS .x. l e a v e s ] |
[l e f t %+JJ .x. l e f t] ]


APPLY DOWN> leave+VBD
left
APPLY UP> leaves
leave+NNS
leave+VBZ
leaf+NNS
```

Morphologie : transducteurs à états finis (FST)

Lexique du français en FST :

- taille une fois compilé : **800 Ko** (français), 500 Ko (anglais)
- *lookup* immédiat
- gestion des langues agglutinantes, des cycles, ...

→ importance des **ressources langagières**

Désambiguïsation : le problème

permis → permettre+V — permis+N :

- un mot = plusieurs analyses
- mots inconnus

→ Comment faire ? Qu'utilise l'humain ?

Désambiguïsation : le problème

permis → permettre+V — permis+N :

- un mot = plusieurs analyses
- mots inconnus

→ Comment faire ? Qu'utilise l'humain ?

Il m'a permis d'habiter chez lui.

Solutions ?

Désambiguïsation : modèles de Markov

Chaîne de Markov [1913] :

- une variable ne dépend que du présent
- la probabilité d'une variable ne dépend que de celle la précédant
  - pas de mémoire
  - homogène (ne dépend pas de la place dans la chaîne)

Désambiguïsation : HMM

- Etant donnée une suite de symboles :
  - Orange [Nom propre, Adjectif, Nom], au cours du [Préposition], matin [interjection, Adverbe, Nom], a [Préposition, Nom], permis [Ppassé, Nom, Adjectif], la [Article, Pronom, Nom], prise [Nom, Ppassé], de [Préposition], bénéfices [Nom]
- On considère la probabilité qu'un symbole se trouve après deux autres symboles
- On prend la plus grande

Tagging : étiquettage morpho-syntaxique

tokenization

+

analyse morphologique + lemmatisation

+

désambiguïsation

Tagging : exemple

Exercice

Montrez les différentes étapes et possibilités de tagging de :
Le chef d'orchestre donne le la.

Lemmatisation *vs stemming*

Le *stemming* c'est la racinisation :

$$\text{imaginait} \rightarrow \text{imagin}$$

La lemmatisation c'est retrouver l'entrée de dictionnaire :

$$\text{imaginait} \rightarrow \text{imaginer}$$

Qualité de l'étiquetage morpho-syntaxique

ou *POS tagging*

Exactitude (*accuracy* en anglais, à ne pas confondre avec la précision) :

- `TreeTagger` (1994) : 95,7 % [Allauzen and Bonneau-Maynard, 2008]
- `MElt` (2010) : près de 98 % [Denis and Sagot, 2010]

Quelle différence concrète ?

Qualité de l'étiquetage morpho-syntaxique

ou *POS tagging*

Exactitude (*accuracy* en anglais, à ne pas confondre avec la précision) :

- `TreeTagger` (1994) : 95,7 % [Allauzen and Bonneau-Maynard, 2008]
- `MElt` (2010) : près de 98 % [Denis and Sagot, 2010]

Quelle différence concrète ?

96 % d'exactitude, environ 10 mots par phrase
→ sur 10 phrases, un mot mal étiqueté dans 4 phrases

98 % d'exactitude → deux fois moins d'erreurs

Analyse syntaxique : le problème

Un énoncé est une suite ordonnée de mots :

- quels mots ?
- **dans quel(s) ordre(s) ?**
- pour quel sens ?
- dans quelles langues ?

Analyse syntaxique : les approches

Deux types d'approches en TAL :

- analyse en constituants
- analyse en dépendances

La constituance (Antoine Gautier)

Une phrase n'est pas une simple concaténation d'unités minimales du type A + B + C + D, qui serait équivalente à B + A + D + C

Observez :

(1) a. Virginie mange une salade.
≠ Une salade mange Virginie.
b. Joey est (stupide et gentil) → Joey est beau.
c. (((Paul) mange) (une salade))

→ La phrase est une hiérarchie et non une **concaténation** d'éléments.

Justifier la notion de constituant (Antoine Gautier)

On observe cependant que certaines suites de mots fonctionnent comme des mots uniques. Autrement dit :

Ils ont la même distribution qu'un mot unique.

(2) a. Le frère de Marie dormait profondément.
b. Paul dormait profondément.

Ils sont déplaçables en bloc, mais pas séparément :

(3) Il dormait profondément, le frère de Marie.

a. *Il de Marie dormait profondément, le frère.
b. *Le frère dormait profondément de Marie.

Ils peuvent constituer des énoncés autonomes :

— Qui dormait profondément ?

— Le frère de Marie.

Ils peuvent être conjoints :

Joey est (bête) → Joey est (bête et gentil)

Têtes et catégories (Antoine Gautier)

La plupart des syntagmes (ou groupes) s'organisent autour d'un mot central qui détermine le reste de sa constitution, la tête du syntagme. La catégorie du syntagme se déduit alors de la catégorie de sa tête.

Nom	SN	[Le frère de Marie] dort profondément.
Préposition	SPrep	J'ai discuté [avec le frère de Marie].
Verbe	SV	Je [connais le frère de Marie].
Adjectif	SAdj	Je suis [très mécontent de mon travail].
Adverbe	SAdv	[Malheureusement pour vous], il va pleuvoir.

Analyse syntaxique : les constituants

Vous connaissez sans doute les catégories suivantes :

- GN, GV, GP (équivalentes à SN, SV, SP)

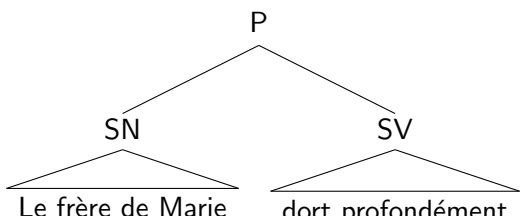
Si je vous demande d'analyser grammaticalement la phrase suivante, vous savez le faire :

Jean aime la belle Marie.

La représentation des constituants (Antoine Gautier)

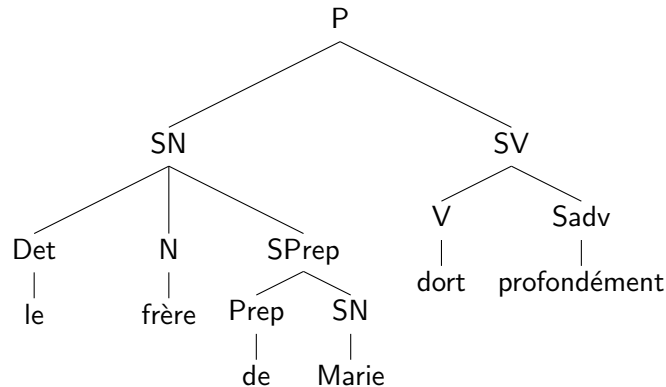
Deux représentations équivalentes :

- Crochets : [SN le frère de Marie] [SV dort profondément].
- Arbre :

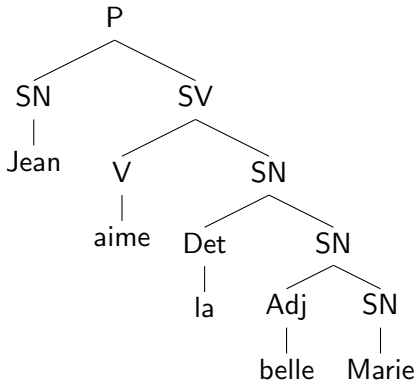


Analyse en Constituants Immédiats (Antoine Gautier)

Faire un découpage systématique de l'énoncé en syntagmes et en mots, jusqu'à ce qu'il ne reste plus que des mots.

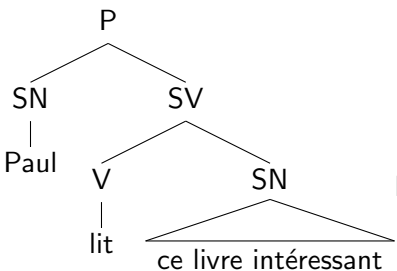


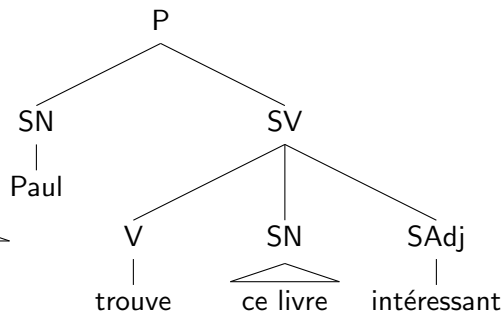
Dans notre cas : *Jean aime la belle Marie*



L'ambiguïté structurale (Antoine Gautier)

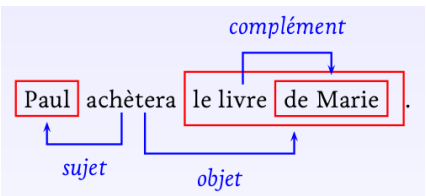
La même séquence de mots peut être un syntagme dans une phrase, mais pas dans une autre.





Les fonctions syntaxiques

Les fonctions syntaxiques sont des relations entre un mot et un constituant qui dépend syntaxiquement de ce mot.



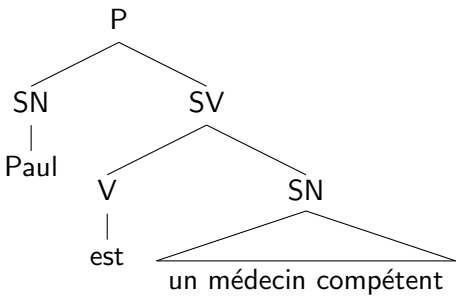
Les différentes fonctions se caractérisent par des propriétés syntaxiques :

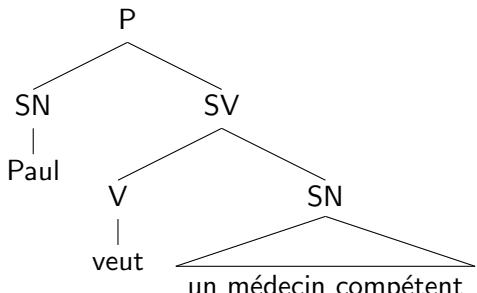
- En français, seul le sujet s'accorde en nombre et en personne avec le verbe.
- En français, seul l'objet direct d'une phrase active peut être transformé en sujet au passif.

Fonctions vs. Positions

(4) **Même position, différentes fonctions**

a. Paul veut un médecin compétent. . . . . *Objet*
b. Paul est un médecin compétent. . . . . *Attribut*





(5) **Même fonction, différentes positions**

a. Je veux un gâteau au chocolat. . . . . *Objet*
b. Que voulez-vous ? . . . . . *Objet*

Analyse syntaxique en dépendances : un exemple

- l'ambiguïté est partout, parfois même pour l'humain
- les principales étapes du TAL
- les solutions trouvées (les principes)
- la loi de Zipf
- les rappels de syntaxe

Allauzen, A. and Bonneau-Maynard, H. (2008).
Training and evaluation of pos taggers on the french multitag corpus.
In Nicoletta Calzolari (Conference Chair), Khalid Choukri, B. M. J. M. J. O. S. P. D. T., editor, Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco. European Language Resources Association (ELRA).
http ://www.lrec-conf.org/proceedings/lrec2008/.

Denis, P. and Sagot, B. (2010).
Exploitation d'une ressource lexicale pour la construction d'un étiqueteur morphosyntaxique état-de-l'art du français.
In Traitement Automatique des Langues Naturelles : TALN 2010, Montréal, Canada.