Corpus, ressources et linguistique outillée · M2SOL034

CM 1 : Introduction, notions de base et aperçu du cours

Ljudmila PETKOVIĆ

Semestre 2, 2024-2025
7 février 2025

Sorbonne Université
Master « Langue et Informatique » (M1 ScLan)
UFR Sociologie et Informatique pour les Sciences Humaines

LETTRES SORBONNE UNIVERSITÉ

Informations pratiques

Administrivia

- contact : ljudmila.petkovic@sorbonne-universite.fr
- Moodle : https://moodle-lettres-24.sorbonne-universite.fr/course/view.php?id=3866
- GitHub : https://github.com/ljpetkovic/M2SOL034

Modalités d'évaluation :

- contrôle continu : rendus en TD (25 %)
- contrôle continu : présentations des étudiant·e·s (25 %)
- examen final (50 %)

Les projets seront à réaliser sur ordinateur.

LETTRES SORBONNE UNIVERSITÉ

Contenu

Syllabus

1. Introduction, notions de base et aperçu du cours
2. Fondamentaux de la textométrie et `TXM`
3. Textométrie avancée
4. Automates finis
5. Reconnaissance d'entités nommées
6. Présentations des étudiant·e·s

Syllabus

1. **Introduction, notions de base et aperçu du cours**
2. Fondamentaux de la textométrie et `TXM`
3. Textométrie avancée
4. Automates finis
5. Reconnaissance d'entités nommées
6. Présentations des étudiant·e·s

Comprendre la relation entre linguistique et informatique

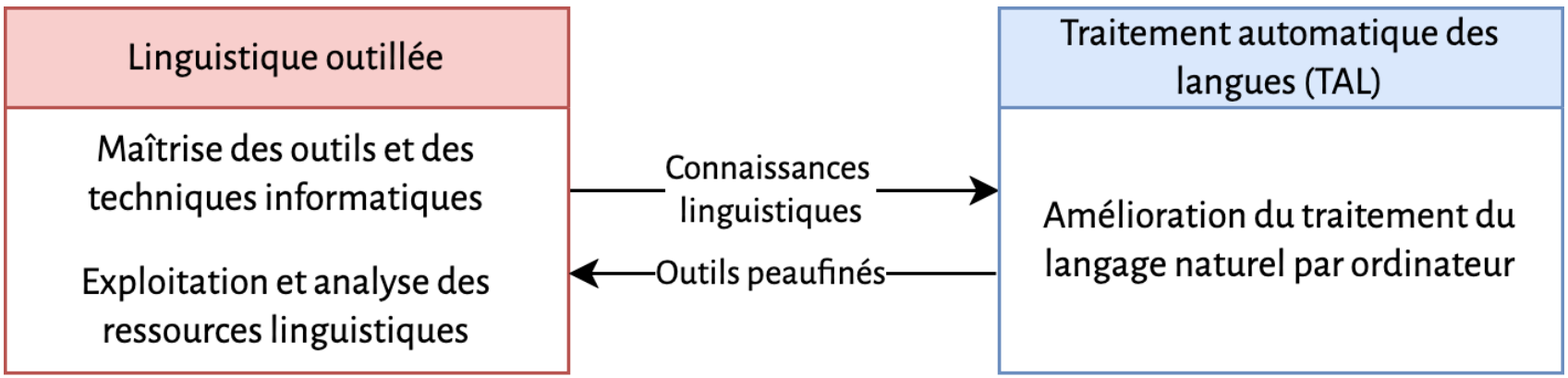

Figure 1 – Adapté de **hodac**.

« *Un corpus ne se lit pas, il s'interroge* »

Corpus électronique

Ensemble de données linguistiques (textes écrits ou retranscriptions de discours oraux) rassemblées en un seul et même endroit, qui peut être interrogé automatiquement via une interface en ligne ou via un programme informatique hors ligne (concordancier).

- échantillon représentatif d'une (variante de) langue
  - variante géographique ; registre, domaine de spécialité...
- peut être associé à d'autres corpus à des fins de comparaison
- données brutes (le corpus ne contient que le texte)
- données annotées (corpus étiqueté).

(**loock2017web**)

LETTRES SORBONNE UNIVERSITÉ

Linguistique outillée

Outils d'analyse linguistique des corpus de textes :

- concordancier
- logiciel de textométrie
- plateforme d'interrogation de corpus

Thématiques centrales :

- Apport des corpus en sciences du langage
- Manipulation de corpus : observation de contextes, analyses quantitatives, textométrie, projection de lexiques et de patrons
- Diversité des corpus
- Études linguistiques outillées d'un corpus diversifié

LETTRES SORBONNE UNIVERSITÉ

TAL

angl. *natural language processing* (NLP)

Thématiques centrales :

- Panorama des techniques et des niveaux d'analyse linguistique
- Applications :
  - OCR [1] : reconnaissance optique de caractères
  - NER [2] : reconnaissance d'entités nommées
  - extraction terminologique
- Programmation (p. ex. en Python)
- Outils d'étiquetage et d'analyse morphosyntaxiques
- Évaluation de l'efficacité des outils du TAL

1. *Optical Character Recognition*
2. *Named Entity Recognition*

LETTRES SORBONNE UNIVERSITÉ

Écosystème

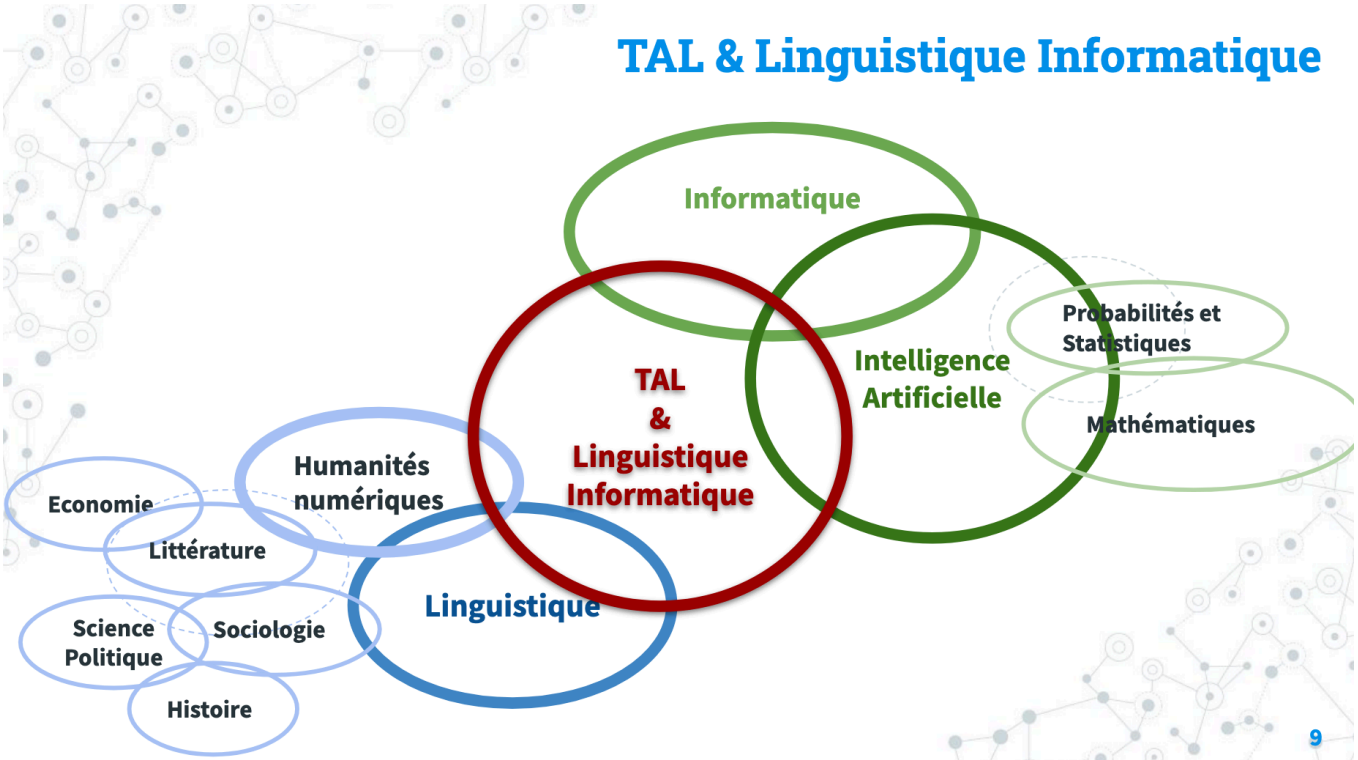

Figure 2 – Le TAL et la linguistique informatique dans l'écosystème des sciences (**boisson**).

Fondamentaux du TAL

Envergure

Domaine de recherche pluridisciplinaire à l'intersection de la linguistique et de l'informatique $\rightarrow$ IA, apprentissage artificiel.

(**poibeau**)

Modéliser et reproduire, à l'aide de machines, la capacité humaine à produire et à comprendre des énoncés linguistiques

(**yvon2010petite**)

- analyse (compréhension) des langues
- génération de texte
- ingénierie linguistique : focus sur les aspects pratiques
- linguistique informatique : focus sur la linguistique

LETTRES SORBONNE UNIVERSITÉ

Applications

Liste non exhaustive :

- traduction automatique
- agents conversationnels
- correcteur orthographique
- prédiction du prochain mot
- classification du texte
- annotation morphosyntaxique

Évolution du domaine

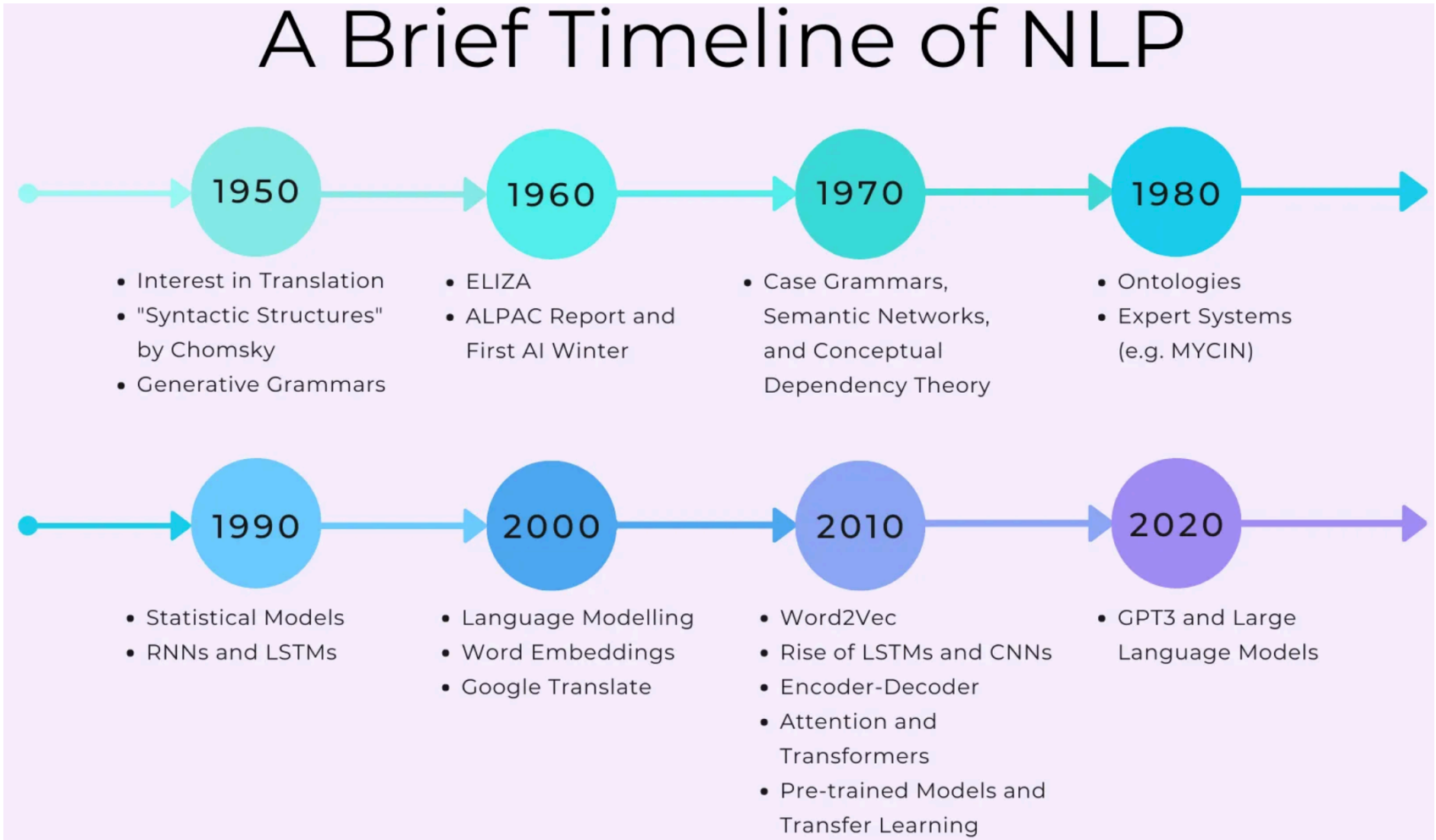

Figure 3 – L'histoire du TAL (**chiusano2022**).

Techniques

Systèmes à base de **règles**

- construites par un humain (linguiste)
- saisie manuelle

Systèmes **orientés données** (angl. *data-driven*)

- apprentissage (non) supervisé
- à partir d'exemples annotés par un humain

LETTRES SORBONNE UNIVERSITÉ

Représentations vectorielles

Plongements (angl. *embeddings*)

- informations linguistiques représentées par des vecteurs de caractères, de mots, de phrases, de paragraphes, de documents
- + : représentations continues, calculs de similarité
- - : ne prend pas en compte la polysémie d'un mot

Limite sémantique des plongements statiques

Chaque mot a exactement une représentation vectorielle fixe.

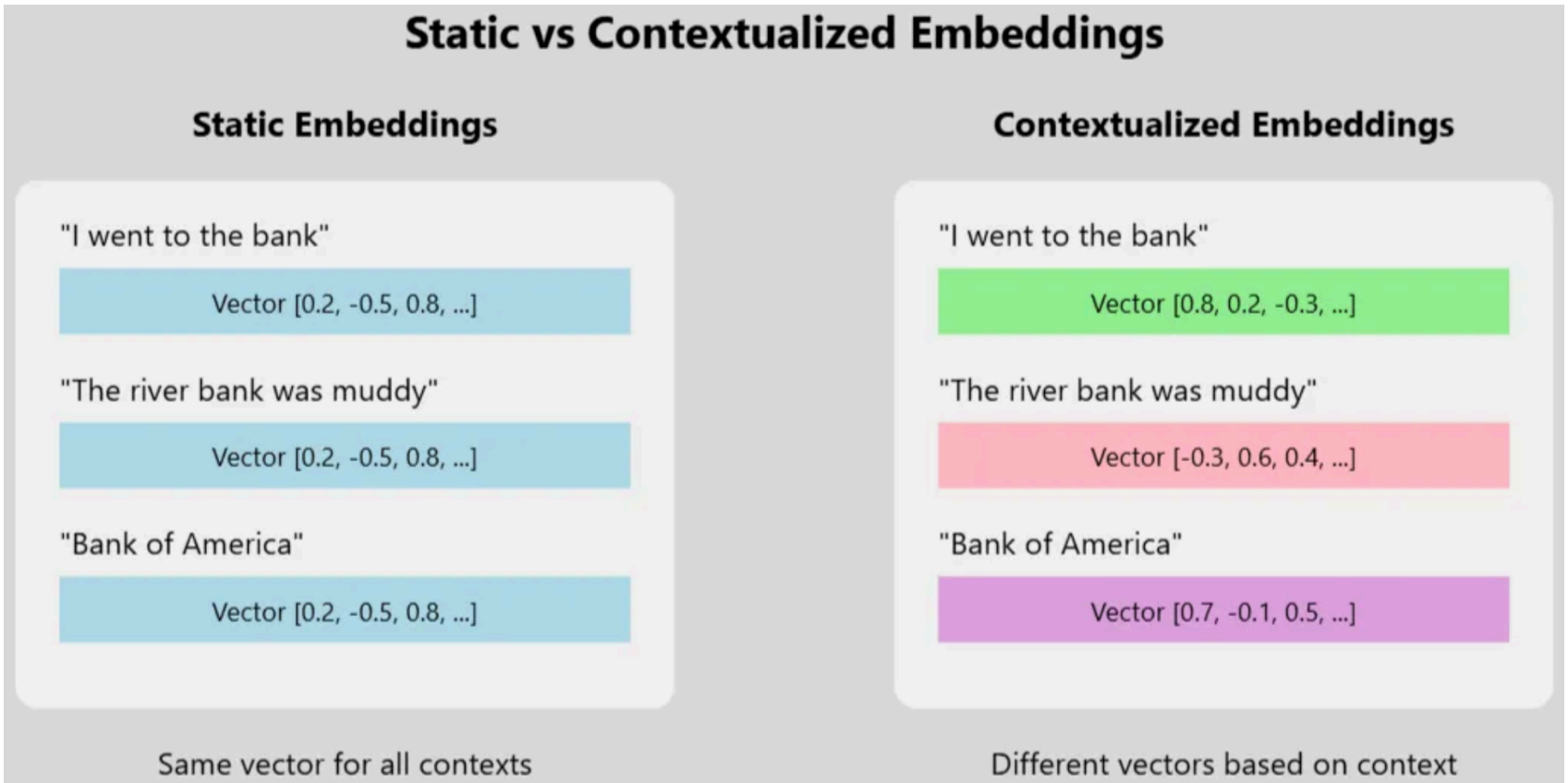

Figure 4 – Plongements statiques et contextuels (**manikanth2024**).

Modèles de langage

- doit distinguer les séquences (non-)grammaticales
- construction d'un modèle de langue statistique permet d'assigner une probabilité à une séquence de mots

(**chomsky1957**)

  - « je vais à la Sorbonne » ✓
  - « je à la Sorbonne vais » ✗

Modèles de langage : apprentissage à partir des données

Entraînement

- non supervisé (pré-entraînement) :
  - apprentissage à partir des grands corpus de textes
- supervisé (*fine-tuning*)
  - annotation des données pour une tâche spécifique
  - spécialisation du modèle
  - le modèle reçoit l'*input* et l'*output* correct
  - il apprend à partir d'exemples
  - p. ex. REN : Marie $\rightarrow$ *Marie* [PER]

Prédiction

Le modèle prédit le résultats sur des données non vues.

LETTRES SORBONNE UNIVERSITÉ

État de l'art : apprentissage profond

angl. ***D**eep **L**earning*

- techniques inspirées par les réseaux de neurones biologiques
- des informations très riches associées aux mots
- modélisation des passages textuels entiers (phrase, §)
- grands modèles de langage : BERT, ...$\mathrm{GPT4}$
- gourmandes en ressources, exigent une puissance de calcul
- introduit des biais : racisme, sexisme etc.
- langues dotées (français) et peu dotées (quechua)

Ambiguïté

Comment traiter informatiquement l'ambiguïté du langage ?

Ambiguïté **lexicale**

- Ambiguïté catégorielle :
  - la ferme (N)
  - je ferme (V)
- Polysémie :
  - roi de France
  - roi de la jungle

Ambiguïté **syntaxique** et **résolution de coréférence**

- Fabrice regarde [Laetitia] [avec son télescope].

Segmentation en mots

Tokenisation

Transformation d'un texte en une série de *tokens* individuels.

Un token représente un mot.

Découpage en tokens (segmentation) :

- *J'ai froid* → `J'` + `ai` + `froid`
- *aujourd'hui* → `aujourd'hui`

Séparateurs : espace, virgule, nouvelle ligne, tab., apostrophes...

Tokenisation

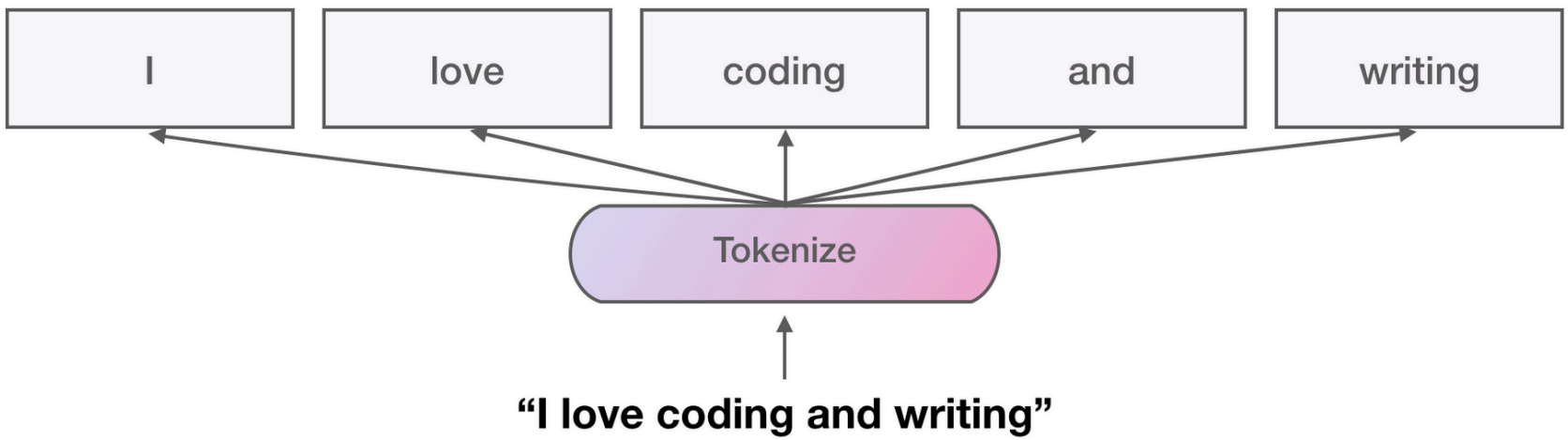

Figure 5 – Exemple de tokenisation (**saravia**).

Segmentation en phrases

- 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

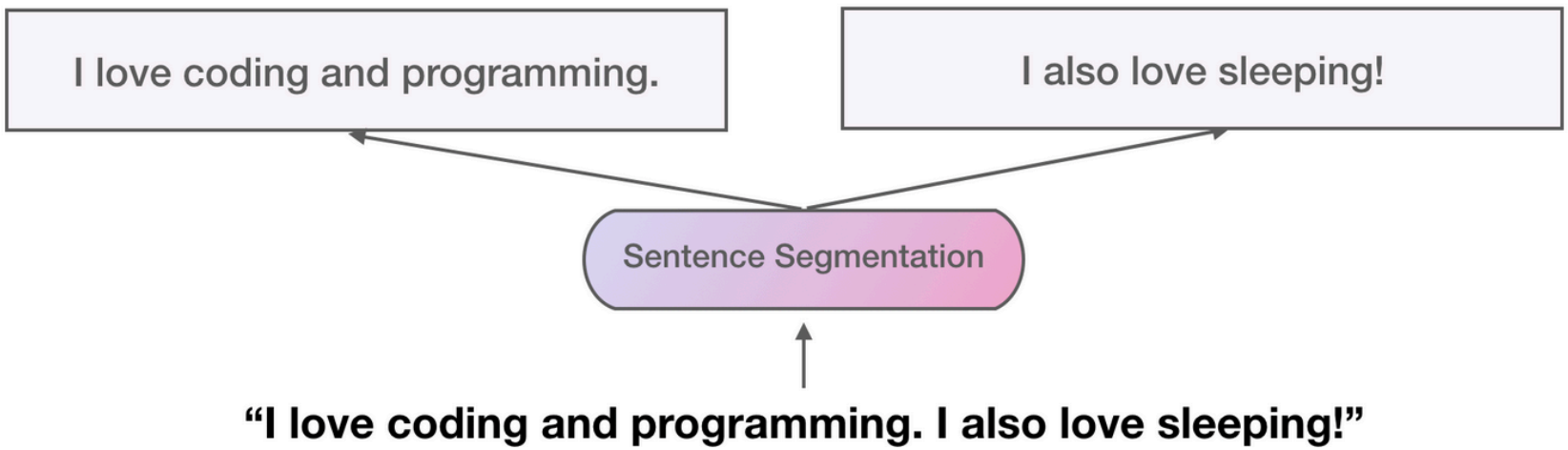

Figure 6 – Exemple de la segmentation en phrases (**saravia**).

Lemmatisation

Réduction des différentes formes d'un mot à une forme canonique : *lemme*.

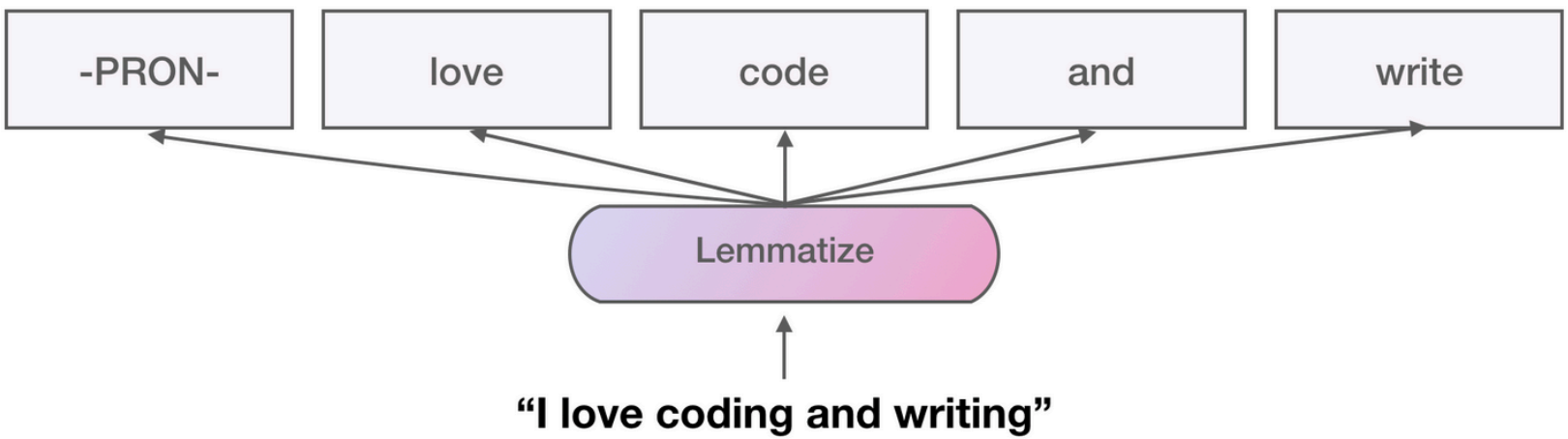

Figure 7 – Exemple de lemmatisation (**saravia**).

Racinisation

Remplacement d'un mot par sa racine morphologique (*radical*).

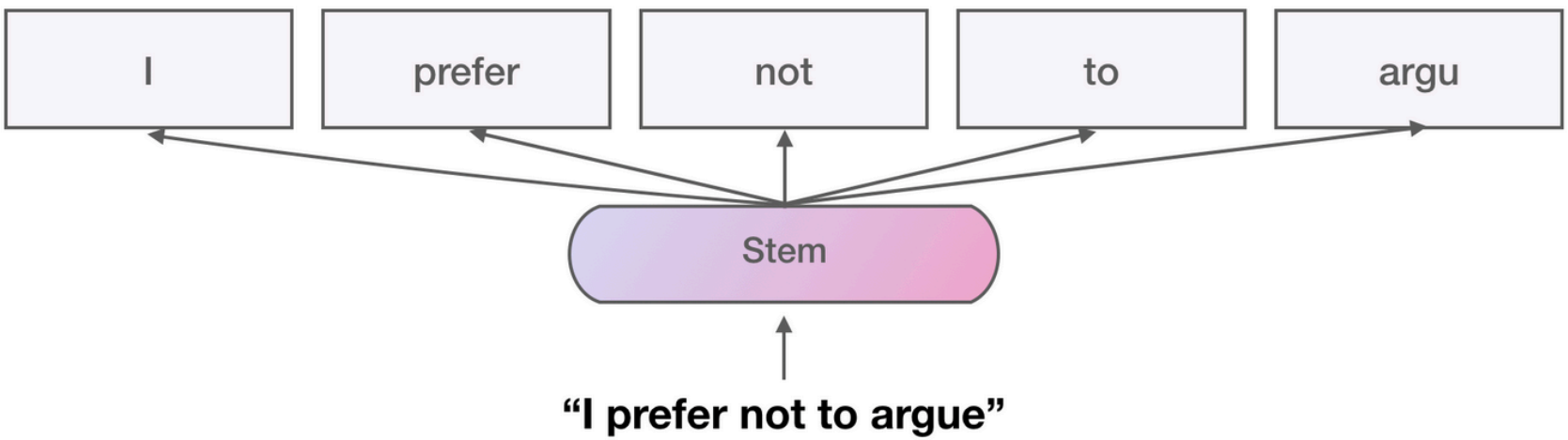

Figure 8 – Exemple de racinisation (**saravia**).

Étiquetage morphosyntaxique

Assigner des informations grammaticales à chaque mot d'un texte.

```
1   Bonne     bon ADJ  _   _   0   _   _   _
2   lecture   lecture  NOUN     _   _   1   _   _   _
3   a   à     ADP  _   _   2   _   _   _
4   bientôt   bientôt  ADV  _   _   3   _   _   _
5   😉   😉    SYM  _   _   4   _   _   _
6   Restez    rester   VERB     _   _   5   _   _   _
7   positif   positif  ADJ  _   _   6   _   _   _
8   ,   ,     PUNCT    _   _   7   _   _   _
9   c'  ce    PRON     _   _   8   _   _   _
10  est  être      AUX  _   _   9   _   _   _
11  primordial   primordial   ADJ  _   _   10  _   _   _
```

Figure 9 – Exemple d'étiquetage morphosyntaxique (**wang2021**).

Évaluation des systèmes de TAL

Précision % d'éléments pertinents détectés par le système parmi tous les éléments détectés

Rappel % d'éléments pertinents détectés par le système parmi tous les éléments à détecter

F-mesure « moyenne harmonique » de la P et du R ; donne le même poids à la P et au R :

$$F\text{-}mesure = \frac{2 \times (P \times R)}{P + R}$$

Exemple de l'évaluation

Corpus de 100 documents, dont 8 pertinents à la requête.

Le système propose 12 documents, dont 6 pertinents.

$$P = \frac{6}{12} = 0,5$$

$$R = \frac{6}{8} = 0,75$$

- « **bruit** » : 6 documents proposés à tort
- « **silence** » : 2 documents « oubliés » qui auraient dû être détectés, mais ne l'ont pas été

Loi de Zipf

Dans un corpus d'une langue, la fréquence d'un mot est inversement proportionnelle à son rang dans la liste globale des mots après le tri par ordre décroissant de fréquence.

Rang	**Mot**	Fréquence
1	de	30
2	des	20
...	...	...
100	optimisation	1

Visualisation de la loi de Zipf

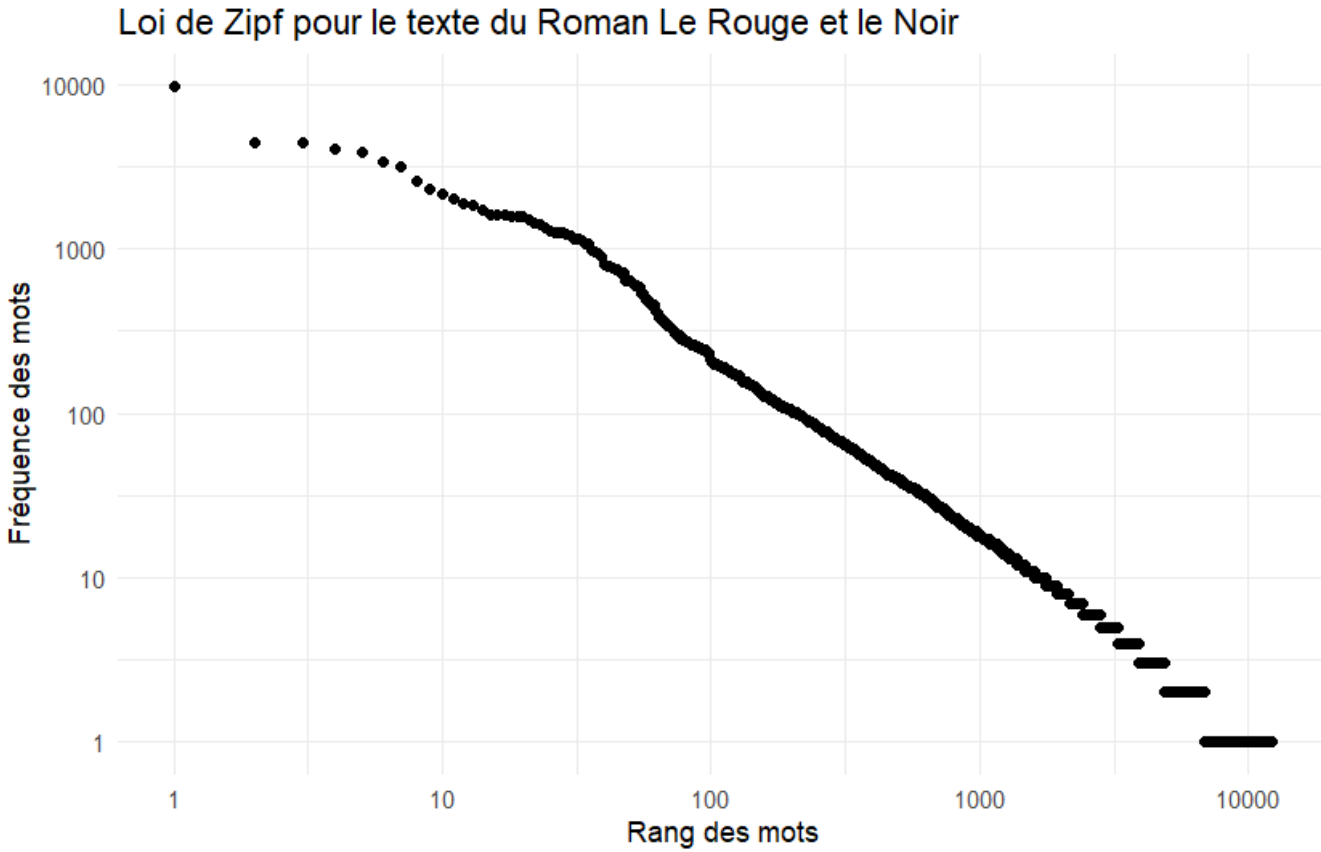

Figure 10 – Exemple de la loi de Zipf (**rherrad**).

Applications de la loi de Zipf dans la fouille de textes

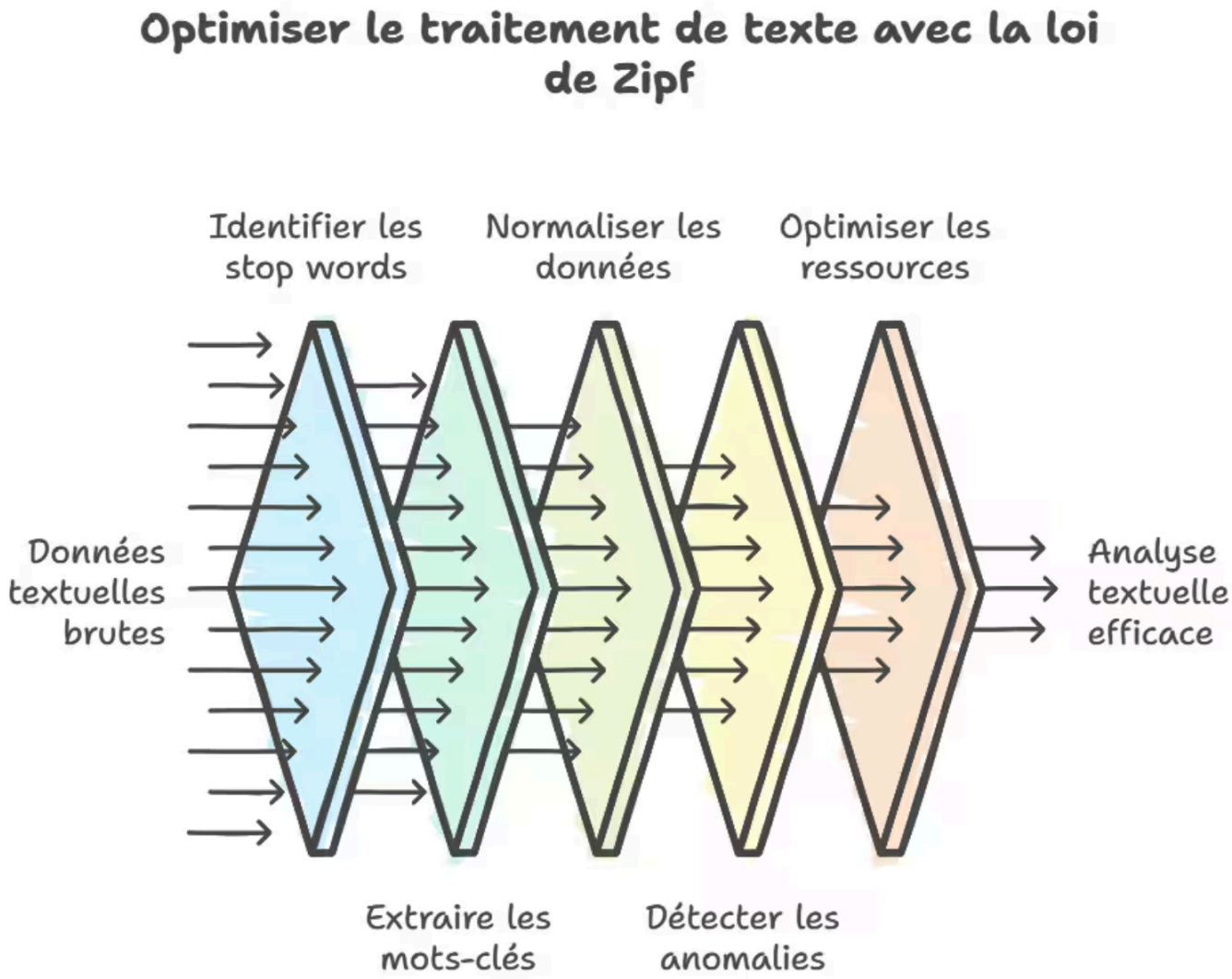

Figure 11 – Optimisation du traitement du texte avec la loi du Zipf (**rherrad**)