

Corpus, ressources et linguistique outillée · M2SOL034

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

Ljudmila PETKOVIĆ

Semestre 2, 2024-2025

30 janvier 2025

Sorbonne Université

Master « Langue et Informatique » (M1 ScLan)

UFR Sociologie et Informatique pour les Sciences Humaines

Informations pratiques

Administrivia

- contact : ljudmila.petkovic@sorbonne-universite.fr
- Moodle : en construction
- 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.

Contenu

Syllabus

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

Comprendre la relation entre linguistique et informatique

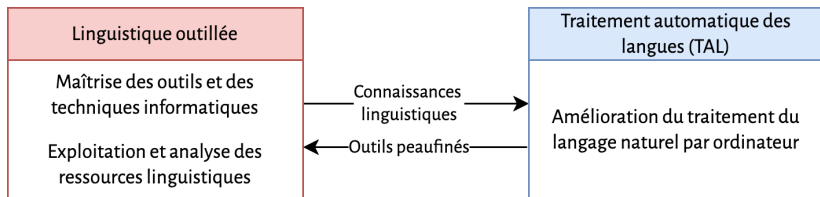


Figure 1 – Adapté de HO-DAC et FABRE (*s.d.*).

« *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).

(LOOCK, 2017)



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

TAL

angl. *natural language processing* (NLP)

Thématiques centrales :

- Panorama des techniques et des niveaux d'analyse linguistique
- Applications :
 - OCR¹ : reconnaissance optique de caractères
 - NER² : 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*

Écosystème

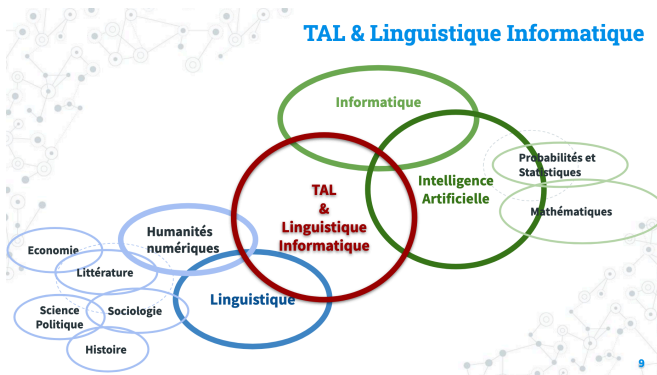


Figure 2 – Le TAL et la linguistique informatique dans l'écosystème des sciences
(BOISSON, 2022).

Fondamentaux du TAL

Envergure

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

(POIBEAU, 2021)

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

(YVON, 2010)

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

Applications

Liste non exhaustive :

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

Évolution du domaine

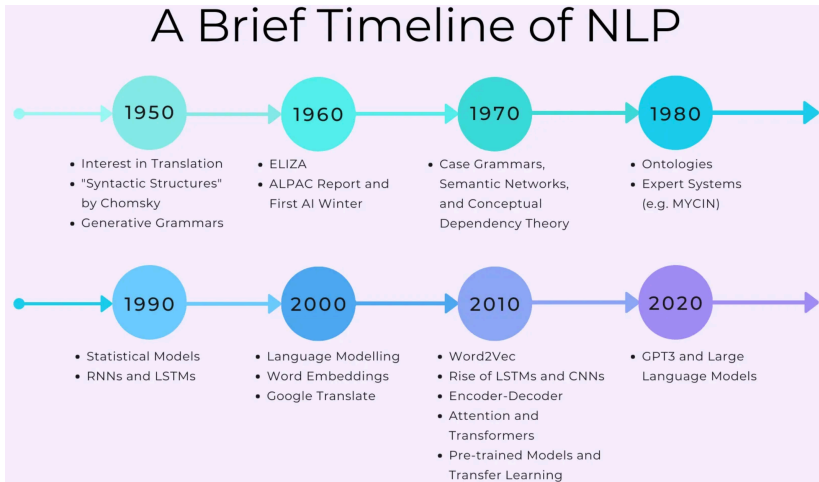


Figure 3 – L'histoire du TAL (CHIUSANO, 2022).

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

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.

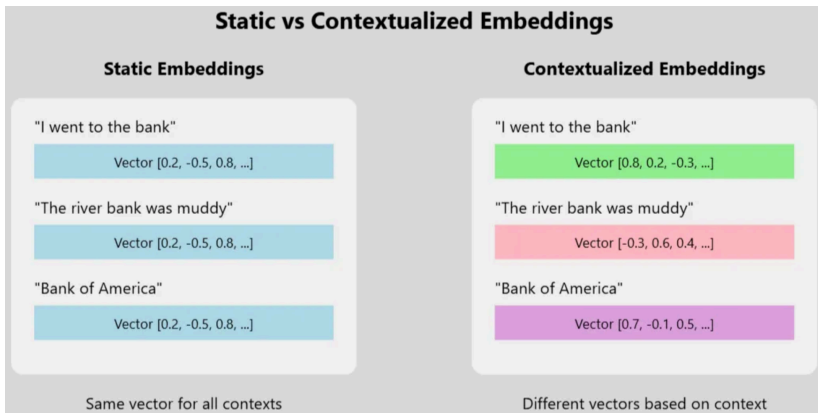


Figure 4 – Plongements statiques et contextuels (MANIKANTH, 2024).

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

(CHOMSKY, 2002)

- « 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 → *Marie* [PER]

Prédiction

Le modèle prédit le résultats sur des données non vues.

État de l'art : apprentissage profond

angl. *Deep Learning*

- 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, ...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

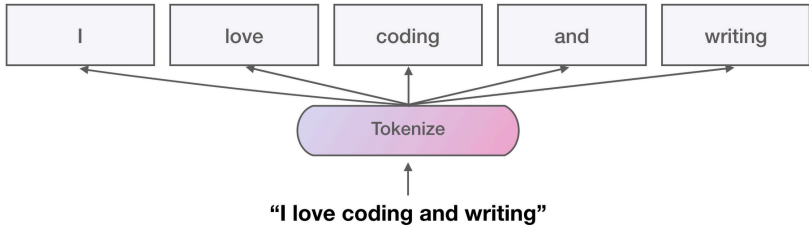


Figure 5 – Exemple de tokenisation ([SARAVIA, s.d.](#)).

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

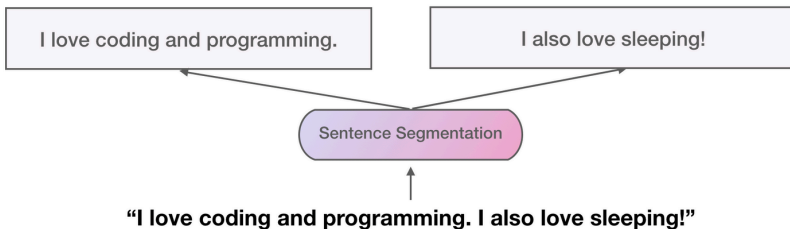


Figure 6 – Exemple de la segmentation en phrases (SARAVIA, *s.d.*).

Lemmatisation

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

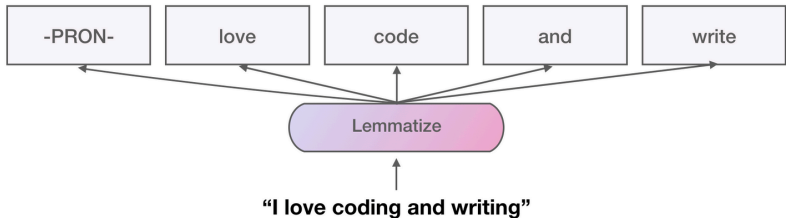


Figure 7 – Exemple de lemmatisation (SARAVIA, *s.d.*).

Racinisation

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

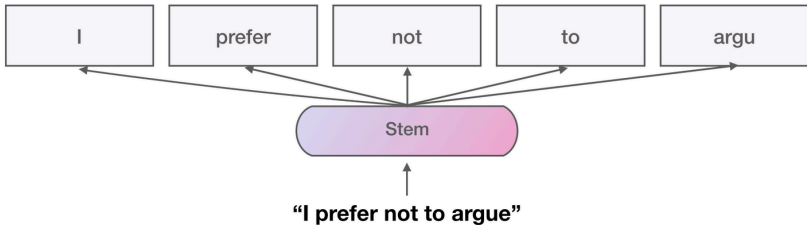


Figure 8 – Exemple de racinisation (SARAVIA, *s.d.*).

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 (WANG, 2021).

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

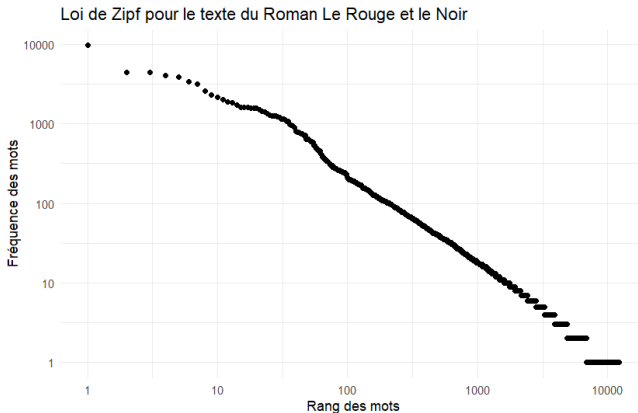


Figure 10 – Exemple de la loi de Zipf ([RHERRAD, s.d.](#)).

Applications de la loi de Zipf dans la fouille de textes

Optimiser le traitement de texte avec la loi de Zipf

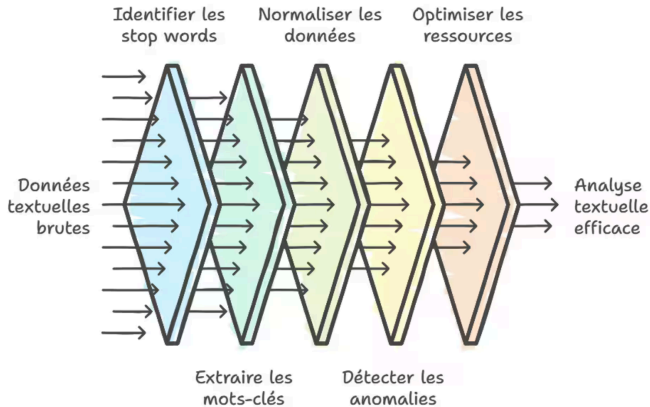


Figure 11 – Optimisation du traitement du texte avec la loi du Zipf (RHERRAL)

Références



BOISSON, J. (2022). ***Comment les ordinateurs interagissent-ils avec nous dans les langues humaines ?*** Les lundis de l'IA de Paul Valéry,
https://edunumrech.hypotheses.org/files/2023/01/JBoisson_lundi_IA_TAL_ac_Paris_dec22.pdf. Consulté le 30 janvier 2025 (*voir p. 11*).



CHIUSANO, F. (2022). ***A Brief Timeline of NLP***. Medium,
<https://medium.com/nlplanet/a-brief-timeline-of-nlp-bc45b640f07d>. Consulté le 30 janvier 2025 (*voir p. 15*).



CHOMSKY, N. (2002). ***Syntactic structures***.
https://tallinzen.net/media/readings/chomsky_syntactic_structures.pdf. Mouton de Gruyter (*voir p. 19*).



HO-DAC, L.-M. et C. FABRE (s.d.). **SL0F601T – Linguistique outillée et Traitement Automatique des Langues**. <https://www.univ-tlse2.fr/accueil/formation-insertion/sl0f601t-linguistique-outillee-et-traitement-automatique-des-langues-1>. Consulté le 30 janvier 2025 (*voir p. 7*).



LOOCK, R. (2017). **Web, corpus, traduction : l'exploitation par les traducteurs des données du web**. In : *Traduire. Revue française de la traduction* 237, p. 23-32. DOI : <https://doi.org/10.4000/traduire.939> (*voir p. 8*).



MANIKANTH (2024). **Understanding Contextualized Word Embeddings : The Evolution of Language Understanding in AI**. Medium
<https://manikanthgoud123.medium.com/understanding-contextualized-word-embeddings-the-evolution-of-language-understanding-in-ai-8bf79a98eb51>. Consulté le 30 janvier 2025 (*voir p. 18*).



POIBEAU, T. (2021). **Traitement Automatique des Langues**. <https://www-universalis-edu-com.bnf.idm.oclc.org/encyclopedie/traitement-automatique-des-langues/>. Consulté le 30 janvier 2025 (*voir p. 13*).



RHERRAD, M. (s.d.). **Loi de Zipf et Text Mining**. Iana-Data, <https://www.iana-data.org/fondamentaux/846-loi-de-zipf-et-text-mining.html>. Consulté le 30 janvier 2025 (*voir pp. 32, 33*).



SARAVIA, E. (s.d.). **Fundamentals of NLP (Chapter 1) : Tokenization, Lemmatization, Stemming, and Sentence Segmentation**. GitHub
https://github.com/dair-ai/nlp_fundamentals/blob/master/1_nlp_basics_tokenization_segmentation.ipynb.
Consulté le 30 janvier 2025 (*voir pp. 24-27*).



WANG, X. (2021). **Comparer trois étiqueteurs morphosyntaxiques (Spacy, StanfordNLP, TreeTagger) sur un corpus formel et un corpus oral.** Medium

<https://xiaoouwang.medium.com/comparer-trois-%C3%A9tiqueteurs-morphosyntaxiques-spacy-stanfordnlp-treetagger-sur-un-corpus-formel-4ca758dde502>. Consulté le 30 janvier 2025 (*voir p. 28*).



YVON, F. (2010). **Une petite introduction au traitement automatique des langues naturelles.** In : *Conference on Knowledge discovery and data mining.*

<https://perso.limsi.fr/anne/coursM2R/intro.pdf>, p. 27-36 (*voir p. 13*).