Au-delà des formes : Part of Speech et Named Entities

Gaël Lejeune (gael.lejeune@sorbonne-universite.fr)

20 novembre 2025

UFR de Sociologie et d'Informatique pour les Sciences Humaines
STIH, Sorbonne Université
CERES (Expérimentations en Méthodes numériques pour les SHS)

Les concepts principaux vus jusqu'ici

- Données, Corpus
- Token
- Expressions régulières
- Classification
- Evaluation : Précision, Rappel, F-mesure

Notions I : autour du corpus

- Ensemble de textes (numériques) réunis pour une étude.
- Corpus littéraires, journalistiques, politiques, etc.
- Notions importantes :
  - **Métadonnées** : auteur, date, source...
  - **Annotation** : information ajoutée (étiquettes, entités...).

Notions II : Tokens, lemmes, types

- **Token** : occurrence d'un mot dans le texte.
- **Type** : forme distincte (vocabulaire).
- **Lemme** : forme canonique (infinitif, masculin singulier...).

Notions II : Tokens, lemmes, types

- **Token** : occurrence d'un mot dans le texte.
- **Type** : forme distincte (vocabulaire).
- **Lemme** : forme canonique (infinitif, masculin singulier...).
- Exemple :
  - ≪ Les enfants jouaient, l'enfant riait ≫ :
  - Tokens : 5
  - Types : *les*, *enfants*, *jouaient*, *l'*, *enfant*, *riait*
  - Lemmes : *le*, *enfant*, *jouer*, *le*, *enfant*, *rire*

Au-delà des formes ?

- Comprendre ce qu'est le **POS tagging** (étiquetage morpho-syntaxique).
- Comprendre ce qu'est la **Reconnaissance d'entités nommées** (NER).
- Découvrir leurs usages en **Humanités Numériques**.
- Se familiariser avec quelques **outils** et **modèles** actuels.

Part of Speech Tagging par l'exemple

Exemple POS : extrait littéraire annoté

≪ Gervaise regardait la pluie tomber, sans rien dire. ≫ [1] *)*

Annotation POS :

```
 Gervaise/PROPN regardait/VERB la/DET pluie/NOM tomber/VERB
,/PUNCT
sans/ADP rien/PRON dire/VERB
./PUNCT
```

Exemple POS : extrait littéraire annoté

≪ Gervaise regardait la pluie tomber, sans rien dire. ≫ [1])

Annotation POS :

```
 Gervaise/PROPN regardait/VERB la/DET pluie/NOM tomber/VERB
,/PUNCT
sans/ADP rien/PRON dire/VERB
./PUNCT
```

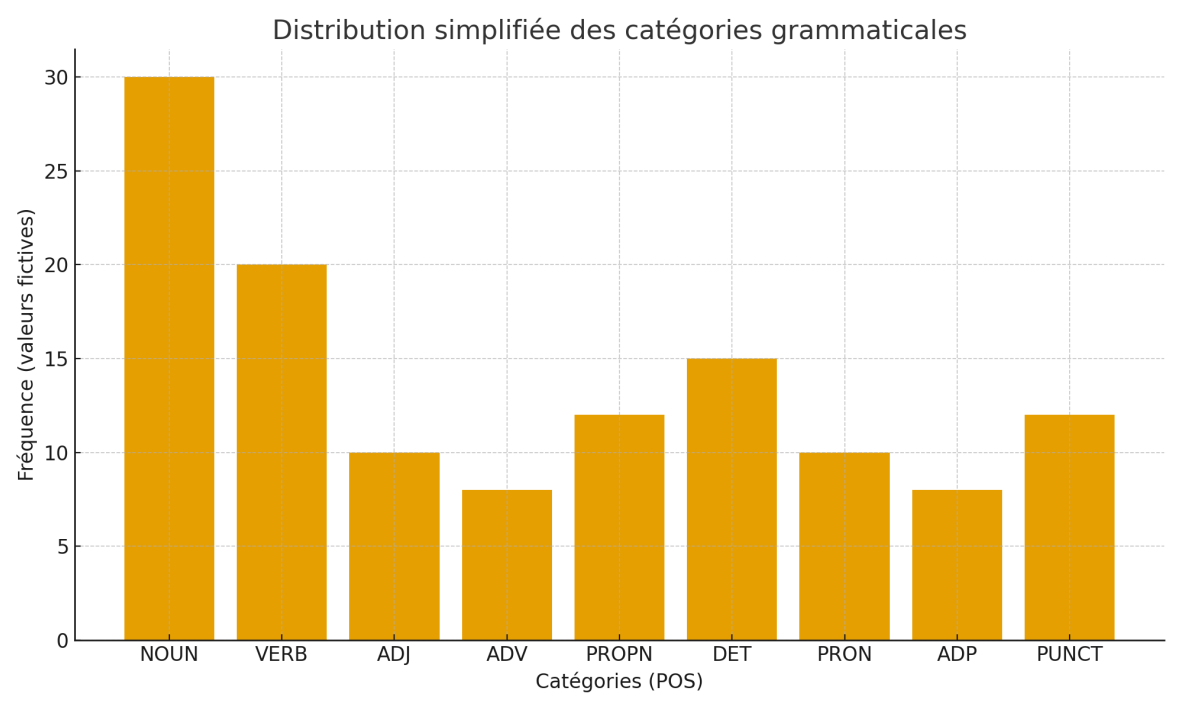


Figure 1 – Distribution des catégories sur un extrait

Catégories grammaticales

- Nom, verbe, adjectif, adverbe, pronom, déterminant ...
- Catégories fines :
  - Verbe transitif / intransitif, nom propre / commun, etc.
- Les jeux d'étiquettes (tagsets) varient selon les projets.

Ambiguïtés linguistiques

- **Ambiguïté lexicale** :
  - ≪ basse ≫ : nom (instrument) ou adjectif (hauteur).
- **Ambiguïté catégorielle** :
  - ≪ livre ≫ : nom ou verbe.
- **Ambiguïté orthographique** :
  - Majuscules, ponctuation, OCR bruité.

Ambiguïtés POS illustrées

- ≪ **livre** ≫ :
  - *nom* : ≪ J'ai acheté un livre. ≫
  - *verbe* : ≪ Je livre le colis. ≫
- ≪ **basse** ≫ :
  - *nom* : ≪ Il joue de la basse. ≫
  - *adjectif* : ≪ Une note trop basse. ≫

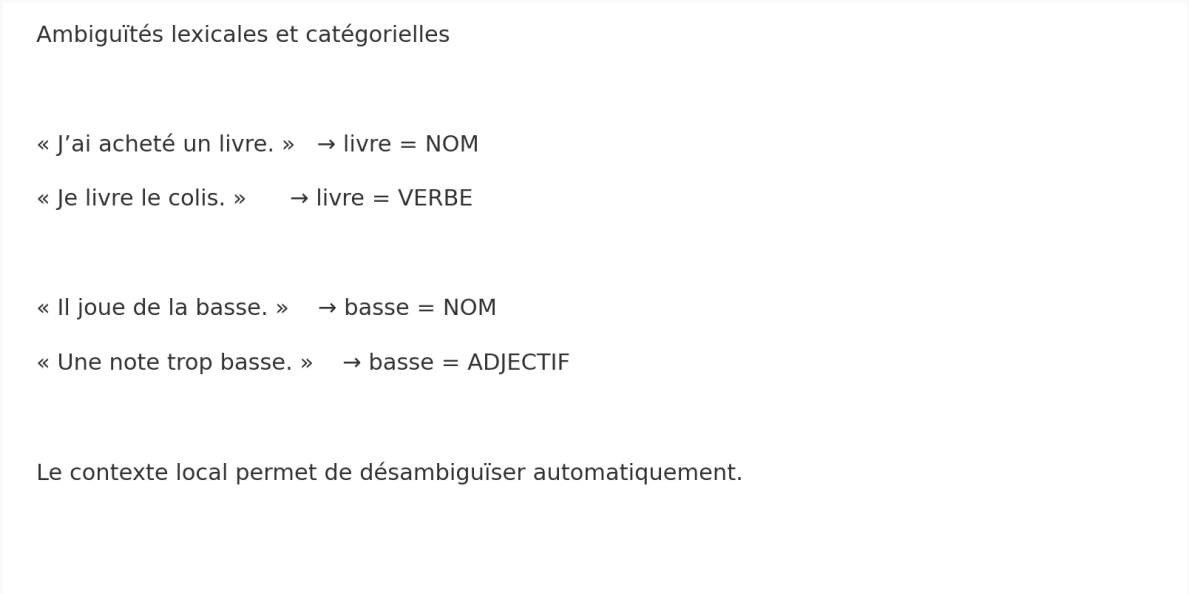


Figure 2 – Ambiguîtés et Contexte

Pourquoi l'étiquetage est difficile ?

- La langue est **variable** (registre, époque, auteur).
- Les contextes peuvent être **rares** ou inhabituels.
- Les corpus des Humanités Numériques :
  - Souvent **bruités** (OCR, orthographe ancienne).
  - Manque de données annotées pour certains genres / périodes.

Pourquoi l'étiquetage est difficile ?

- La langue est **variable** (registre, époque, auteur).
- Les contextes peuvent être **rares** ou inhabituels.
- Les corpus des Humanités Numériques :
  - Souvent **bruités** (OCR, orthographe ancienne).
  - Manque de données annotées pour certains genres / périodes.

Solution : bien définir pour bien réaliser, et bien évaluer !

POS Tagging : notions et méthodes

Définition du POS Tagging

Part-of-Speech (POS) Tagging

- Associer à chaque token une **catégorie grammaticale**.

Exemple :

- « Paris est magnifique » :
- Paris/PROPN est/VER magnifique/ADJ
- ou Paris/NOM est/VER :PRES magnifique/ADJMs . . . ?

Définition du POS Tagging

Part-of-Speech (POS) Tagging

- Associer à chaque token une **catégorie grammaticale**.

Exemple :

- « Paris est magnifique » :
- Paris/PROPN est/VER magnifique/ADJ
- ou Paris/NOM est/VER :PRES magnifique/ADJMs . . . ?

Jeux d'étiquettes (tagsets)

- **Penn Treebank** (anglais).
- **Universal Dependencies** (UD) :
  - Jeu d'étiquettes "**universel**" (ADJ, NOM, VERB, ADV...).
  - Adapté à de nombreuses langues.
- Tagsets spécifiques au français dans certains corpus.

Enjeux du POS Tagging

- Aide à :
  - L'analyse syntaxique.
  - L'extraction d'information.
  - La stylométrie (profil grammatical).
- En Humanités Numériques :
  - Comparaison d'auteurs.
  - Étude de genres (roman, théâtre, presse...).

Approches pour le POS Tagging

- Approches à base de **règles**.
- Approches **statistiques** (HMM, CRF...).
- Approches **neuronales** (LSTM, Transformers...).
- Tendance actuelle : modèles **pré-entraînés**.

Erreurs typiques en POS Tagging

- Confusion nom/adjectif nom/verbe verbe/adjectif.
- Noms propres mal reconnus.
- Structures inhabituelles (poésie, style télégraphique).
- Phrases incomplètes, titres, légendes.

Méthodes de POS Tagging

Approches à base de règles

- Systèmes historiques :
  - Listes de règles linguistiques (si contexte X alors étiquette Y).
- Avantages :
  - Interprétables, contrôlables.
- Inconvénients :
  - Coût de développement élevé.
  - Difficiles à maintenir à grande échelle.

Approches à base de règles : illustration

Idée générale

- Étape 1 : étiquetage initial par **lexique** (mots fréquents, ambiguïtés).
- Étape 2 : **règles de correction** basées sur le contexte.

Approches à base de règles : illustration

Idée générale

- Étape 1 : étiquetage initial par **lexique** (mots fréquents, ambiguïtés).
- Étape 2 : **règles de correction** basées sur le contexte.

Le chat noir mange la souris.

Le/DET/PRO chat/NOM noir/ADJ/NOM mange/VER la/DET souris/NOM/VER ./PUNCT

Règles (style Brill)

- si un mot est DET/PRO et précède un NOM, alors → DET.

  Le → DET

Approches à base de règles : illustration

Idée générale

- Étape 1 : étiquetage initial par **lexique** (mots fréquents, ambiguïtés).
- Étape 2 : **règles de correction** basées sur le contexte.

Le chat noir mange la souris.

Le/DET/PRO chat/NOM noir/ADJ/NOM mange/VER la/DET souris/NOM/VER ./PUNCT

Règles (style Brill)

- si un mot est DET/PRO et précède un NOM, alors → DET.
  Le → DET
- si un mot est ADJ/NOM et suit un NOM, alors → ADJ.

Approches à base de règles : illustration

Idée générale

- Étape 1 : étiquetage initial par **lexique** (mots fréquents, ambiguïtés).
- Étape 2 : **règles de correction** basées sur le contexte.

Le chat noir mange la souris.

Le/DET/PRO chat/NOM noir/ADJ/NOM mange/VER la/DET souris/NOM/VER ./PUNCT

Règles (style Brill)

- si un mot est DET/PRO et précède un NOM, alors → DET.
  Le → DET
- si un mot est ADJ/NOM et suit un NOM, alors → ADJ.
  noir → ADJ
- si un mot est NOM/VER et suit un DET, alors → NOM.
  souris → NOM

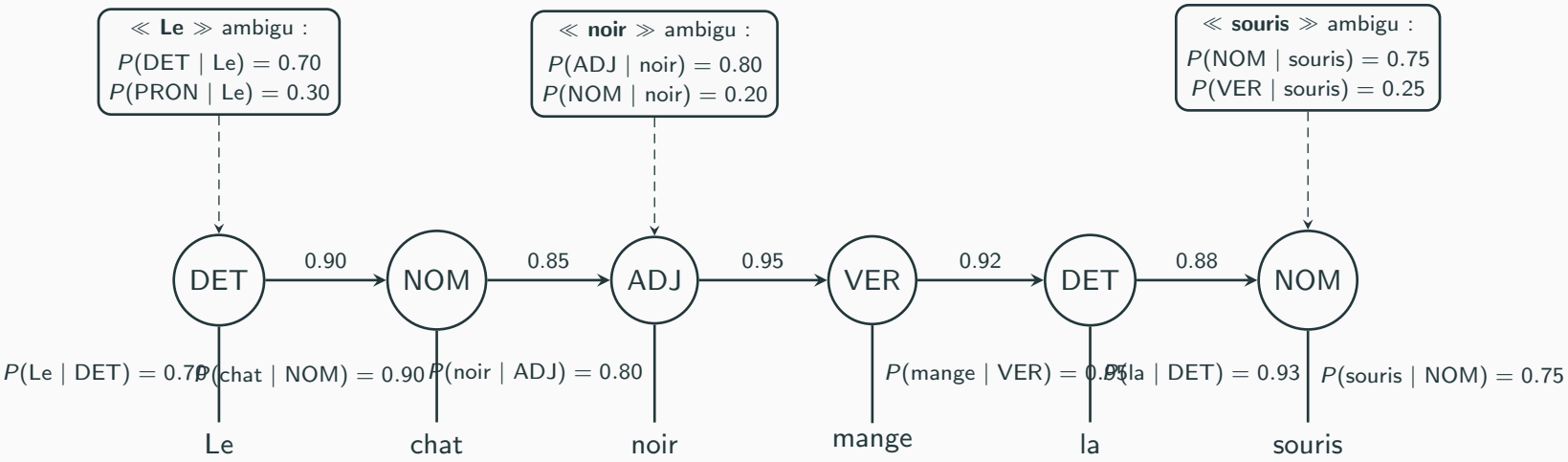
Résultat final :

Le/DET chat/NOM **noir/ADJ** mange/VERB la/DET **souris/NOM** ./PUNCT

Modèles probabilistes (HMM, CRF)

- **Hidden Markov Models (HMM)** :
  - Les étiquettes forment une chaîne cachée.
  - On calcule la séquence la plus probable.
- **Conditional Random Fields (CRF)** :
  - Modélisation discriminante de la séquence d'étiquettes.
- Très utilisés avant l'essor des réseaux de neurones.

Modèle HMM : illustration



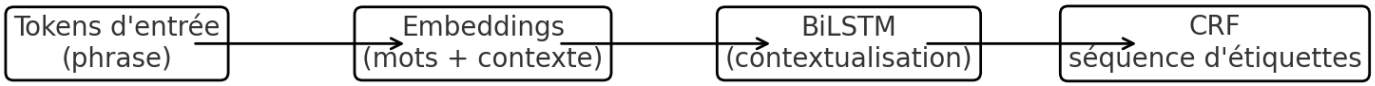
Horizontalemnt : probabilités de transition entre étiquettes
Encadrés : étiquettes possibles pour les mots ambigus.

POS Tagging neuronaux

- **Réseaux récurrents** (LSTM, BiLSTM).
- **BiLSTM + CRF** : architecture classique.
- Avantages :
  - Meilleure prise en compte du **contexte**.
  - Performances supérieures sur de nombreux corpus.

Architecture BiLSTM-CRF

Architecture typique BiLSTM-CRF pour POS/NER



Représentations distribuées (embeddings)

- **Word2Vec**, **GloVe**, **FastText** :
  - Représentent les mots par des **vecteurs**.
- Idée : les mots qui apparaissent dans des contextes similaires ont des vecteurs proches.
- Intégrés dans de nombreux modèles de POS/NER.

Transformers et modèles pour le français

- Architecture **Transformer** (BERT et dérivés).
- Pour le français :
  - **CamemBERT**, **FlauBERT**, etc.
- Fine-tuning possible pour POS + NER.

Évaluation du POS Tagging

- **Exactitude** (accuracy) :
  - proportion de tokens correctement étiquetés.
- **Précision, rappel, F1** pour des étiquettes spécifiques.
- Importance de **corpus de référence** annotés à la main.

Exemples d'erreurs et de limites

- Poésie, style archaïque, néologismes.
- Erreurs d'OCR : mots inconnus ou mal segmentés.
- Changement de langue (citations en anglais, latin, etc.).

Limites pratiques du POS Tagging

- Langues anciennes, orthographe variable.
- Genres peu représentés dans les données d'entraînement.
- Besoin de **réentraînement** ou d'**adaptation** de modèles.

Introduction à la NER

Qu'est-ce que la NER ?

- NER = **Named Entity Recognition**.
- Tâche : détecter et typer des **entités nommées**.
- Exemples de types :
  - Personnes, lieux, organisations.
  - Dates, événements, œuvres.

NER vs POS

- POS : catégorie grammaticale de chaque mot.
- NER : catégorie "sémantique" de certains mots.
- Exemple :
  - « Victor Hugo est né à Besançon » :
  - Victor Hugo = PER, Besançon = LOC
- Les deux tâches sont souvent **complémentaires**.

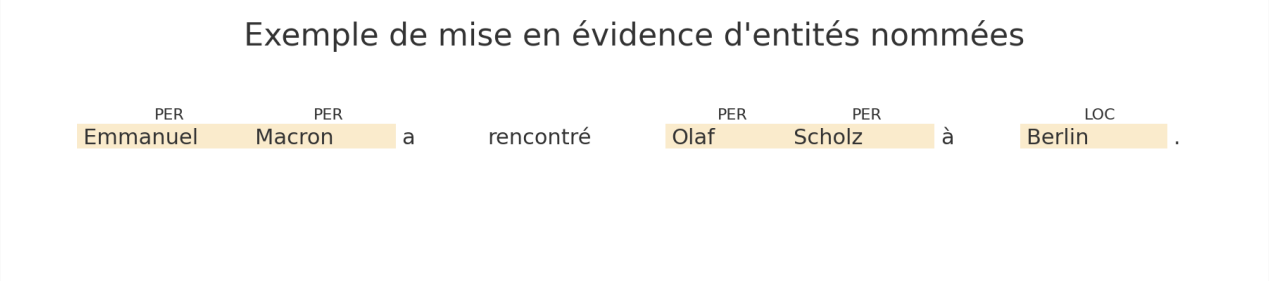
Exemple NER : article de presse

Extrait (Le Monde, 2024) :

Emmanuel Macron a rencontré Olaf Scholz à Berlin . . .

Annotation NER (BIO) :

```
 B-PER Emmanuel
I-PER Macron
O a
O rencontré
B-PER Olaf
I-PER Scholz
O à
B-LOC Berlin
```



Frontières d'entités : difficulté illustrée

- « **Université Paris Cité** » → ORG ou plusieurs entités ?
- « **Ministère de l'Économie et des Finances** » → entité longue.
- « **Louis XIV** » vs « **le roi Soleil** » (référence indirecte).

Jeux d'étiquettes pour la NER

- Jeux standards (CoNLL, etc.) :
  - PER, LOC, ORG, MISC...
- Pour les Humanités :
  - Possibilité de définir des types spécifiques :
  - *Œuvres*, *Personnages fictifs*, *Institutions historiques*...
  - Mais aussi Dates (DATE) !

Ambiguïtés en NER

- **Homonymes** :
  - « Paris » : ville ou personne.
- **Noms communs devenus noms propres** :
  - « Amazon », « Orange », etc.
- **Frontières d'entités** :
  - « Université Paris Cité » : une seule entité ou plusieurs ?

Applications de la NER en Humanités Numériques

- **Cartographie** :
  - géolocalisation des lieux dans un corpus.
- **Prosopographie** :
  - étude systématique des personnages (réels ou fictifs).
- **Réseaux** :
  - qui est associé à qui ? (co-occurrences d'entités).

Méthodes pour la NER

Règles, lexiques et listes d'autorité

- Utilisation de :
  - dictionnaires de noms propres,
  - listes d'autorités (bibliothèques, archives).
- Liens avec le **Linked Open Data** (LOD) :
  - Wikidata, VIAF, etc.
- Avantage : cohérence, connexion à des identifiants stables.

NER statistique (CRF, HMM)

- Approches séquentielles proches du POS.
- NER formulé comme une tâche d'**étiquetage de séquence**.
- Besoin de corpus annotés :
  - textes avec entités marquées (BIO, IOB2...).

NER neuronale (BiLSTM-CRF)

- Architecture très répandue :
  - embeddings de mots + BiLSTM + couche CRF.
- Avantages :
  - bonne prise en compte du contexte,
  - performances élevées sur de nombreux benchmarks.

NER avec Transformers

- BERT, CamemBERT, FlauBERT, etc.
- Principe :
  - modèle pré-entraîné sur de grands corpus,
  - **fine-tuning** pour la tâche NER.
- Avantages :
  - très bonnes performances,
  - prise en compte du contexte global de la phrase.

Évaluation de la NER

- Utilisation de :
  - **Précision, rappel, F1-score** au niveau des entités.
- Problèmes :
  - entités partiellement correctes,
  - types d'entités mal définis.
- Importance de **protocoles d'annotation** clairs.

Cas complexes pour la NER

- **Noms composés** :
  - ≪ Ministère de l'Éducation nationale ≫.
- **Œuvres** :
  - titres longs, ponctuation interne.
- **Entités historiques** :
  - noms obsolètes, changements de frontières, etc.

Comparer des outils NER

- Exemple : `spaCy`, `Flair`, modèles `HuggingFace`.
- Appliquer plusieurs outils sur le même texte.
- Comparer :
  - types d'entités proposés,
  - qualité des annotations,
  - facilité d'usage (licence, documentation, etc.).

Limites de la NER en Humanités Numériques

- Corpus historiques, OCR bruité.
- Références implicites (pronominalisation, ellipses).
- Besoin d'adapter les modèles à des **genres spécifiques**.

Conclusion et ressources

Résumé

- POS tagging :
  - base pour l'analyse morpho-syntaxique.
- NER :
  - extraction d'acteurs, lieux, institutions essentiels.
- En Humanités Numériques :
  - nombreuses applications (cartographie, réseaux,...)
- Défis :
  - qualité des corpus,
  - adaptation aux époques, genres, langues.

Ressources et outils

- Bibliothèques :
  - `spaCy`, `Stanza`, `Flair`, `transformers` (HuggingFace).
- Corpus annotés (selon la langue) :
  - Universal Dependencies, corpus CoNLL, etc.
- Votre travail pour la prochaine fois :
  - tester le code pour la NER dispo sur le moodle
  - être capable de vous ens ervir en cours pour la prochaine fois