Extraction d'Entités Nommées au format BIO avec spaCy dans différentes versions du texte

Weronika BIEGOWSKA 21211975

weronika.biegowska@etu.sorbonne-universite.fr

Date: 04/04/2025

LETTRES
SORBONNE
UNIVERSITÉ

Problématique

- La reconnaissance des entités nommées (REN) et son importance dans le traitement automatique du langage (TAL).
- La la reconnaissance optique de caractères (OCR)

Constitution des Corpus

Tab.1: Corpus

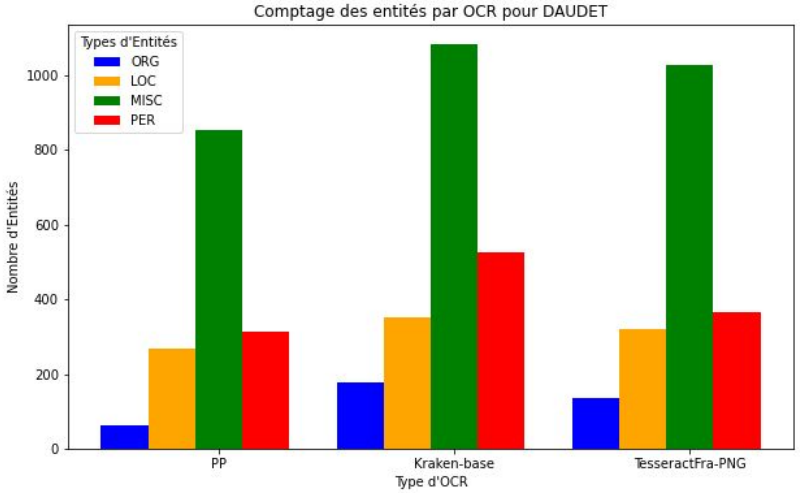
Auteur	Nom du fichier de référence	Nom du fichier de OCR: Kraken	Nom du fichier de OCR: Tesseract
DASH	DASH_chateau-de-Pinon-V1_PP	DASH_chateau-de-Pinon-V1_Kraken-base	DASH_chateau-de-Pinon-V1_TesseractFra-PNG
DAUDET	DAUDET_petit-chose_PP	DAUDET_petit-chose_Kraken-base	DAUDET_petit-chose_TesseractFra-PNG
FLAUBERT	FLAUBERT_education-sentimentale_PP	FLAUBERT_education-sentimentale_Kraken-base	FLAUBERT_education-sentimentale_TesseractFra-PNG
MAUPASSANT	MAUPASSANT_une-vie_PP	MAUPASSANT_une-vie_Kraken-base	MAUPASSANT_une-vie_TesseractFra-PNG
NOAILLES	NOAILLES_la-nouvelle-esperance_PP	NOAILLES_la-nouvelle-esperance_Kraken-base	NOAILLES_la-nouvelle-esperance_TesseractFra-PNG

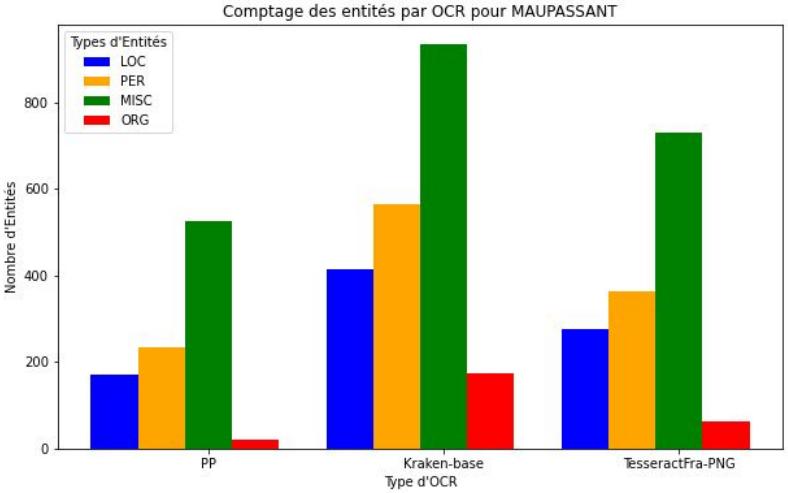
OCR utilisés : **Tesseract** (open-source, généraliste) et **Kraken** (spécialisé pour les documents historiques)

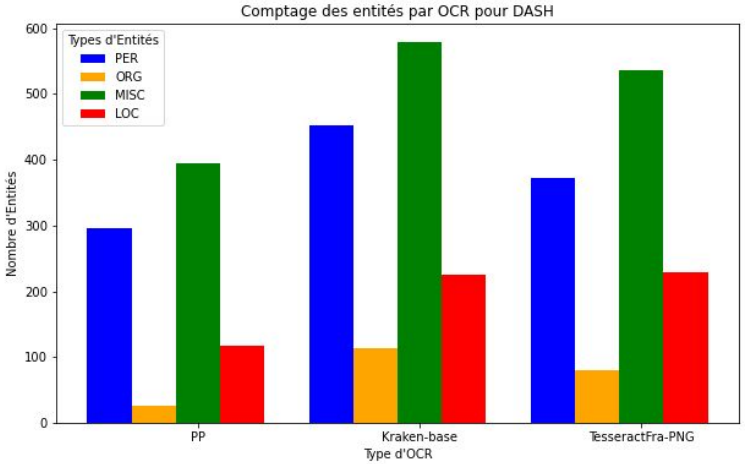
Méthodes d'analyse

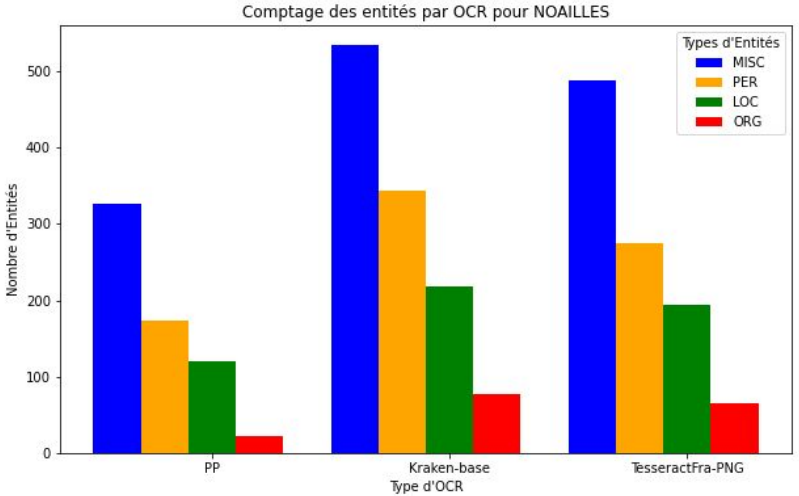
- Utilisation de **spaCy** pour la reconnaissance d'entités.
- Format **BIO** pour l'annotation des entités.
- Reconnaissance d'entités et **POS tagging** (étiquetage de la partie du discours).
- Récupération des tokens avec le label "**PROPN**" (Proper Noun).

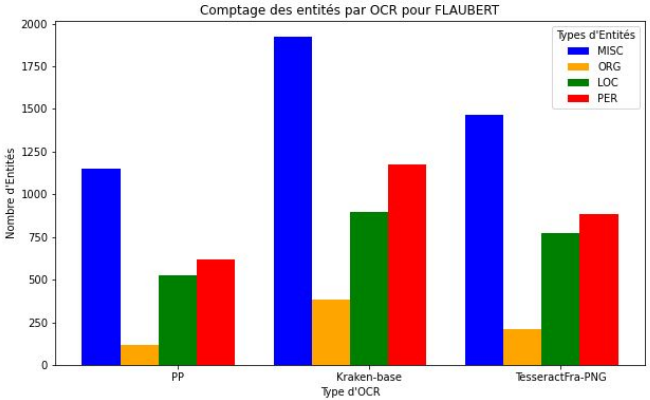
La proportion d’entités pour chaque label sémantique selon les différentes versions des textes.







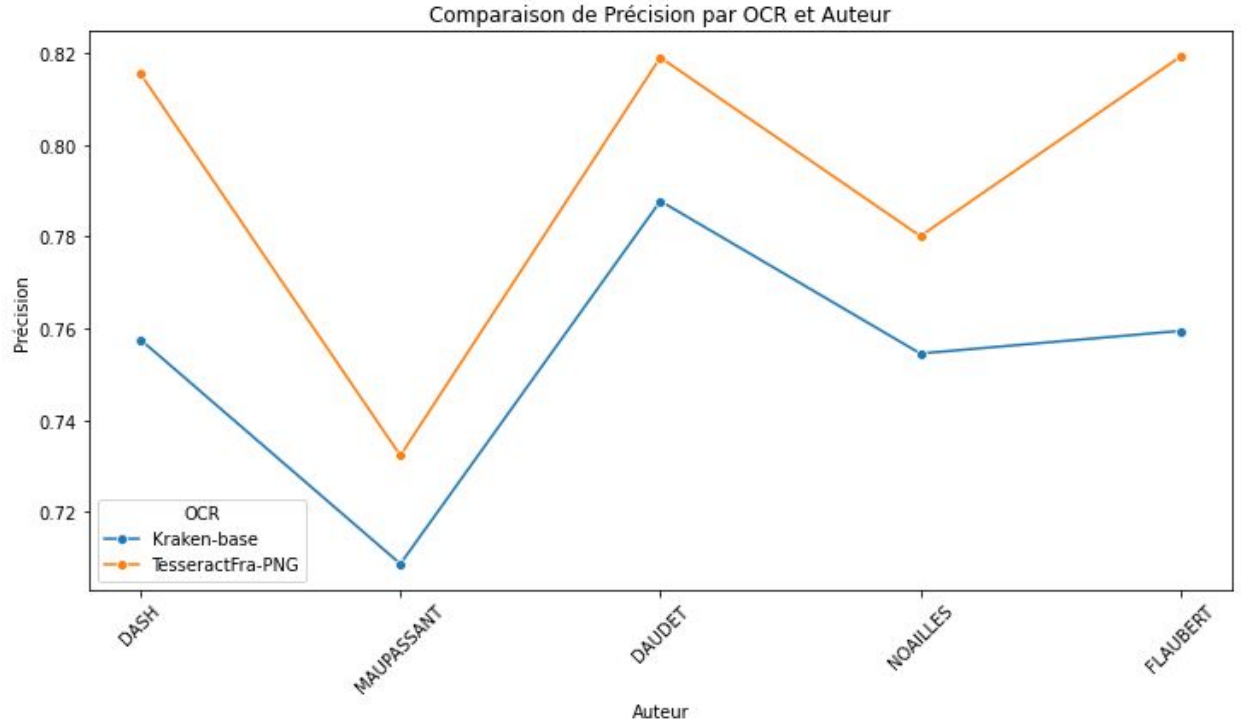




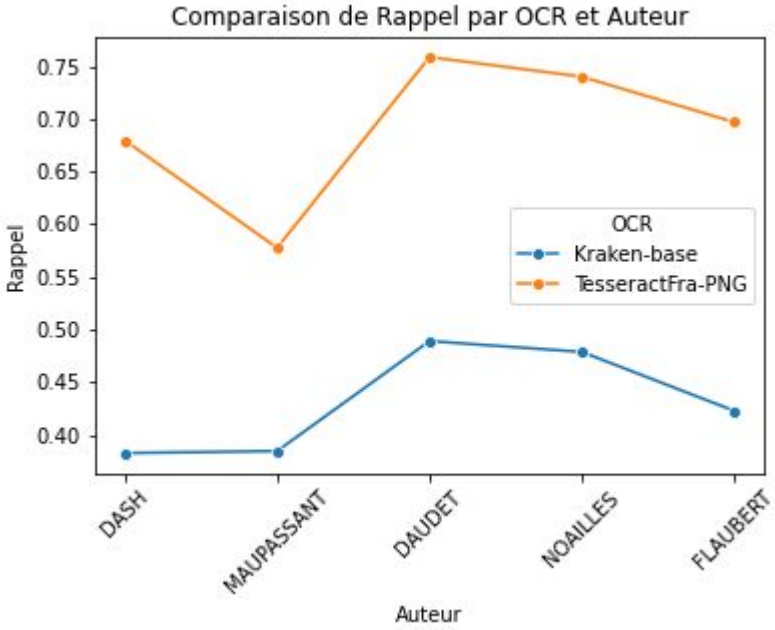
Résultats clés :

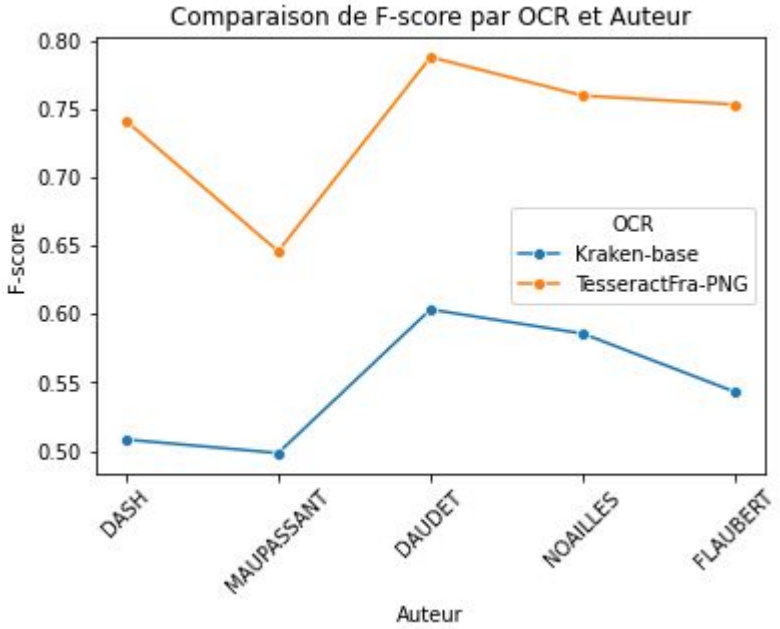
- Plus d'entités nommées ont été retrouvées dans les textes OCR que dans la version de référence.
- Le label sémantique "MISC" est le plus fréquent dans tous les textes et toutes les versions.

Précision

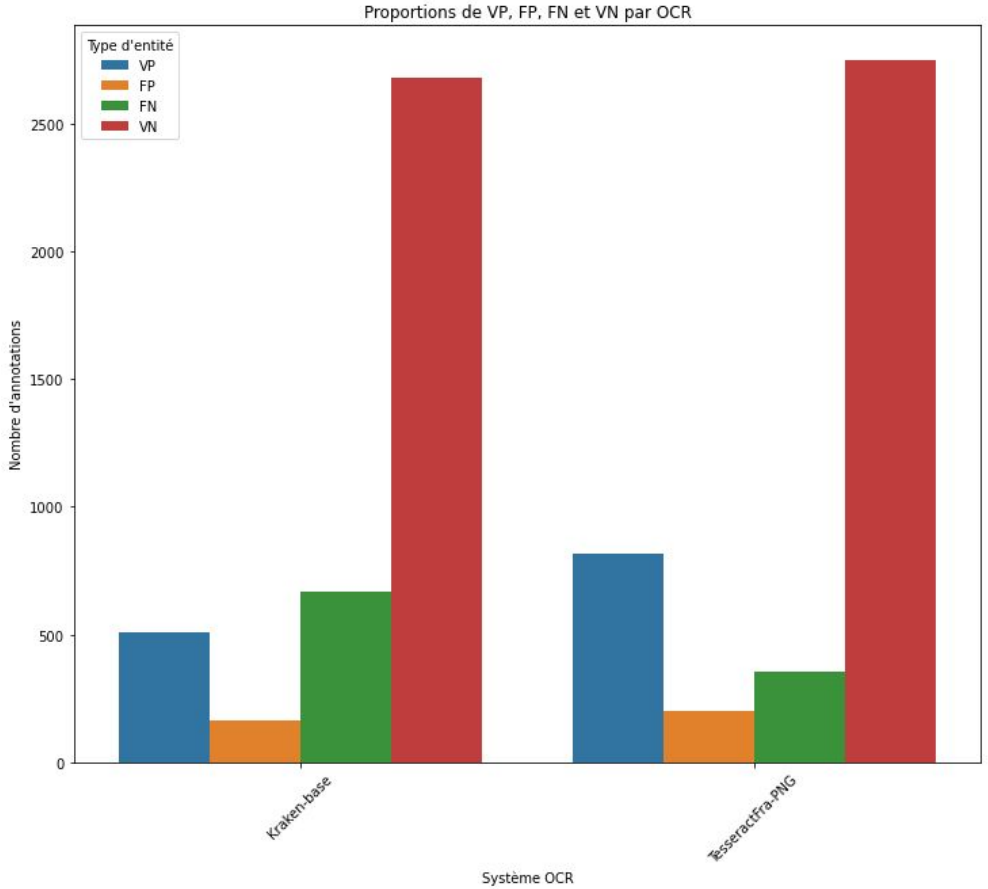

Comparaison de Précision par OCR et Auteur
Précision
0.82
0.80
0.78
0.76
0.74
0.72
OCR
Kraken-base
TesseractFra-PNG
DASH
MAUPASSANT
DAUDET
NOAILLES
FLAUBERT
Auteur

Rappel et F-score

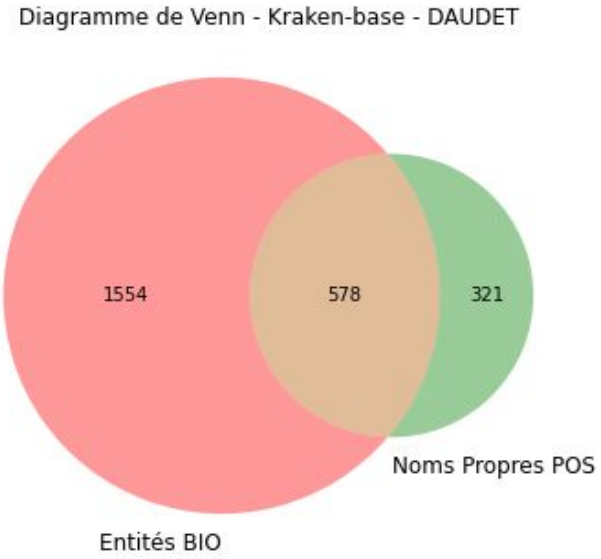


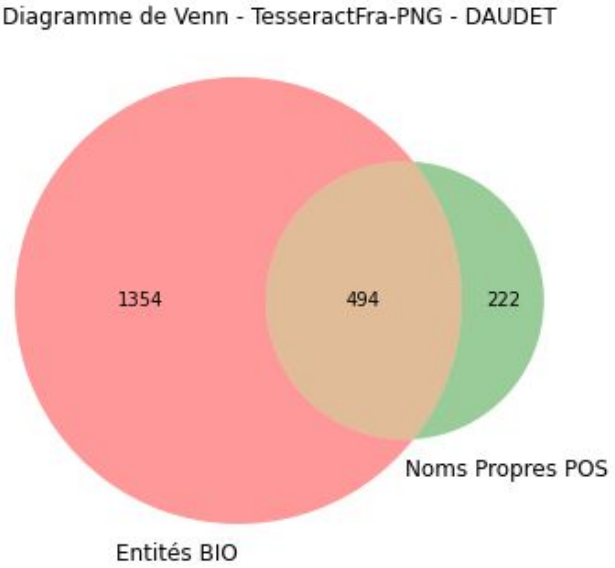


Proportions de VP, FP, VN, FN



Les intersections entre les sorties de REN et les sorties de POS tagging.





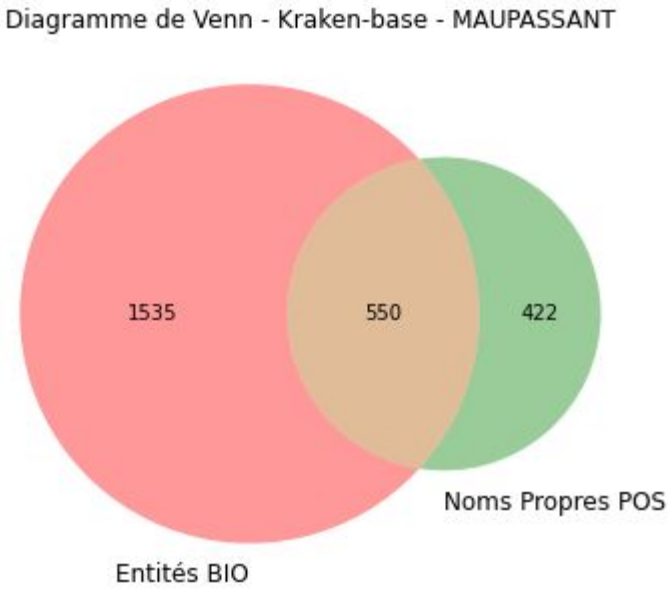


Diagramme de Venn - Kraken-base - MAUPASSANT
1535
550
422
Noms Propres POS
Entités BIO

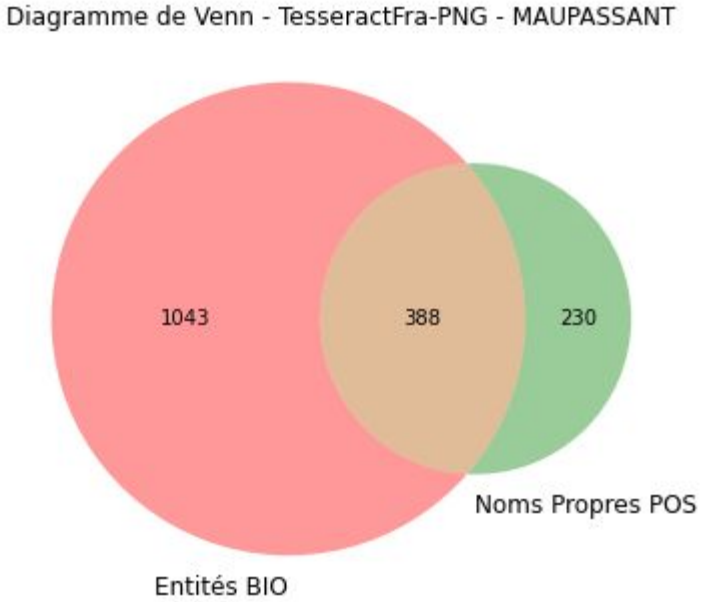


Diagramme de Venn - TesseractFra-PNG - MAUPASSANT
1043
388
230
Noms Propres POS
Entités BIO

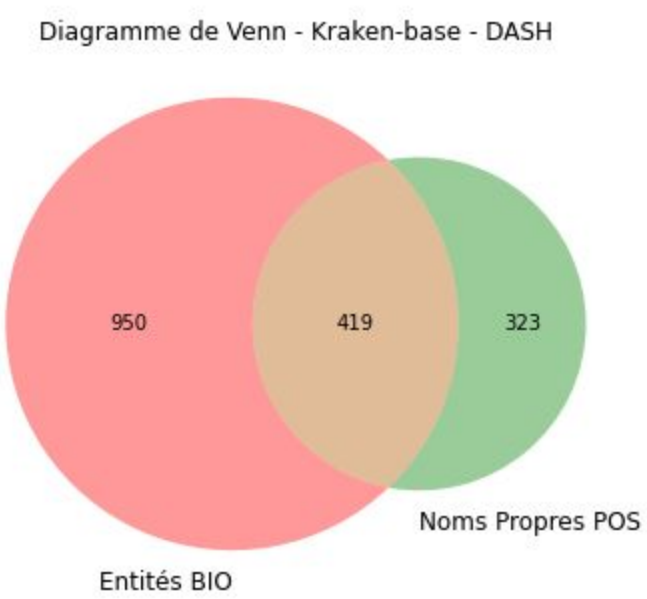


Diagramme de Venn - Kraken-base - DASH
950
419
323
Noms Propres POS
Entités BIO

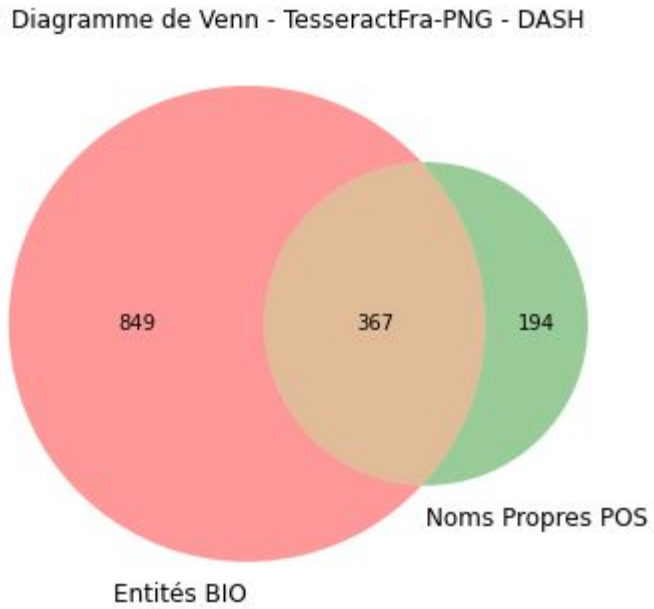


Diagramme de Venn - TesseractFra-PNG - DASH
849
367
194
Noms Propres POS
Entités BIO

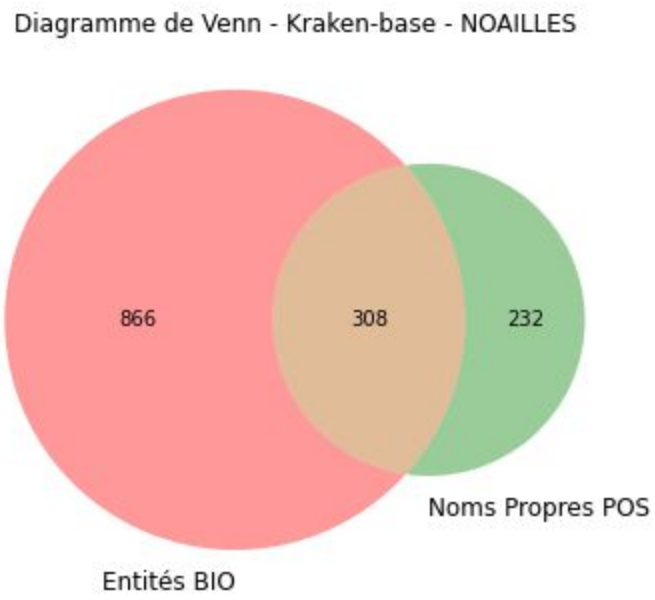


Diagramme de Venn - Kraken-base - NOAILLES
866
308
232
Noms Propres POS
Entités BIO

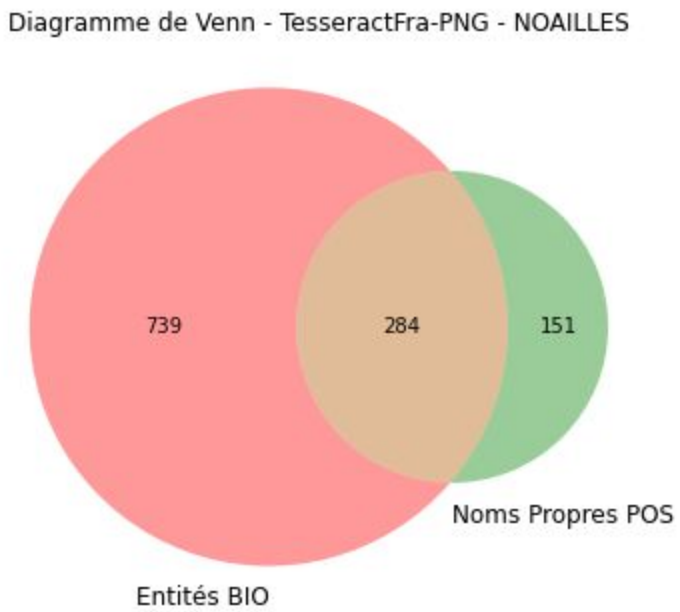


Diagramme de Venn - TesseractFra-PNG - NOAILLES
739
284
151
Noms Propres POS
Entités BIO

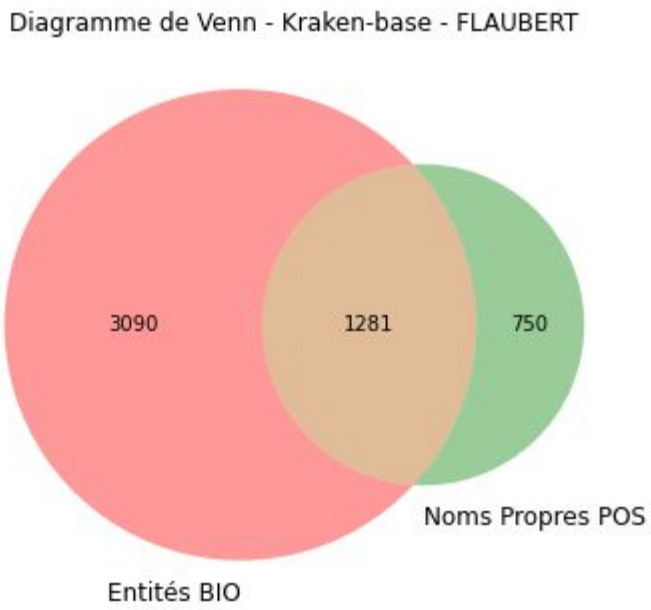


Diagramme de Venn - Kraken-base - FLAUBERT
3090
1281
750
Noms Propres POS
Entités BIO

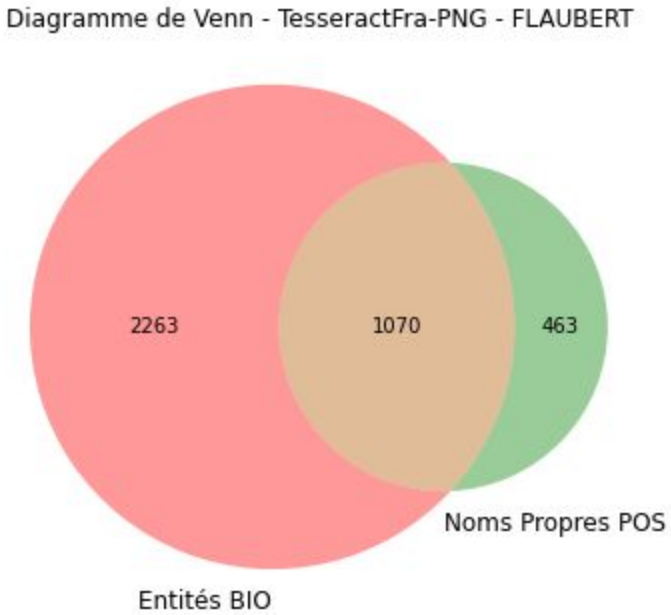


Diagramme de Venn - TesseractFra-PNG - FLAUBERT
2263
1070
463
Noms Propres POS
Entités BIO

Conclusions

- Cette étude a exploré l’extraction d’entités nommées au format BIO avec spaCy sur différentes versions d’un texte, y compris des versions OCR.
-
- Les résultats ont montré que la qualité du texte influence fortement la précision du REN, les erreurs OCR perturbant la détection des entités.
-
- L’analyse via POS tagging a mis en évidence certaines confusions entre entités nommées et noms propres.

Perspectives

- l’utilisation de modèles entraînés sur des textes OCR bruyés, ainsi que des post-traitements linguistiques ou automatiques, pourrait réduire ces erreurs.
- Une comparaison avec des modèles transformers comme BERT pourrait également offrir des résultats plus précis.
- Ce travail souligne les défis du REN sur des textes de qualité variable et ouvre des perspectives pour des approches plus robustes.

Récapitulation

L'objectif: Cette étude a examiné l'extraction d'entités nommées (REN) avec spaCy sur des textes originaux et des versions issues de la reconnaissance optique de caractères (OCR).

Principaux résultats:

- La qualité du texte est un facteur déterminant de la précision de la REN.
- Les erreurs introduites par l'OCR ont un impact négatif sur l'identification des entités nommées.
- Tesseract-Fra-PNG a démontré une meilleure performance globale comparé à Kraken-base.
- L'analyse morphosyntaxique (POS tagging) a révélé des chevauchements et des distinctions entre les entités nommées et les noms propres.

Des stratégies d'amélioration, telles que l'entraînement sur des données OCR bruitées, peuvent améliorer la robustesse des modèles.