DÉSAMBIGUÏSATION DES MOTS POLYSÉMIQUES DE LA VILLE DANS DES ROMANS DE SCIENCE-FICTION

Auteurs:

Sami GUEMBOUR (Univ Gustave Eiffel, ENSG, IGN, LASTIG)

Catherine DOMINGUÈS (Univ Gustave Eiffel, ENSG, IGN, LASTIG)

Bruxelles - 25 juin 2024

JADT 2024 : 17th International Conference on Statistical Analysis of Textual Data

Contexte

- Ce papier s'inscrit dans le cadre du projet PARVIS[1], pour PARoles de VIlleS.

- Le projet PARVIS vise à analyser les représentations de la ville future afin de mettre en lumière les thèmes et les défis associés aux imaginaires urbains futuristes.

- Cet article décrit une tâche de désambiguïsation, annexe d'un travail antérieur [Guembour et al., 2023] réalisé dans le cadre du projet PARVIS.

1 : https://parvis.hypotheses.org/

Contexte

- Dans [Guembour et al., 2023]:

  - Caractérisation de la ville du futur à partir d'un corpus de romans de science-fiction (Corpus PARVIS) en identifiant les éléments urbains, objets et lieux.

  - Utilisation d'une ressource terminologique regroupant des mots de la ville.

  - Analyse des romans où la ville constitue un contexte essentiel, identifiés comme ceux où les mots de la ressource terminologique sont les plus présents.

Présentation du corpus PARVIS

- Le corpus PARVIS regroupe 131 romans de science-fiction, totalisant **1 056 287** phrases et **29 038 420** mots (segmentations effectuées avec : **NLTK** [Bird et al., 2009]).

- Tous les romans du corpus PARVIS sont en français, soit traduits, soit écrits directement dans cette langue.

- Les romans du corpus ont été publiés entre 1961 et 2020.

Construction d'une ressource terminologique

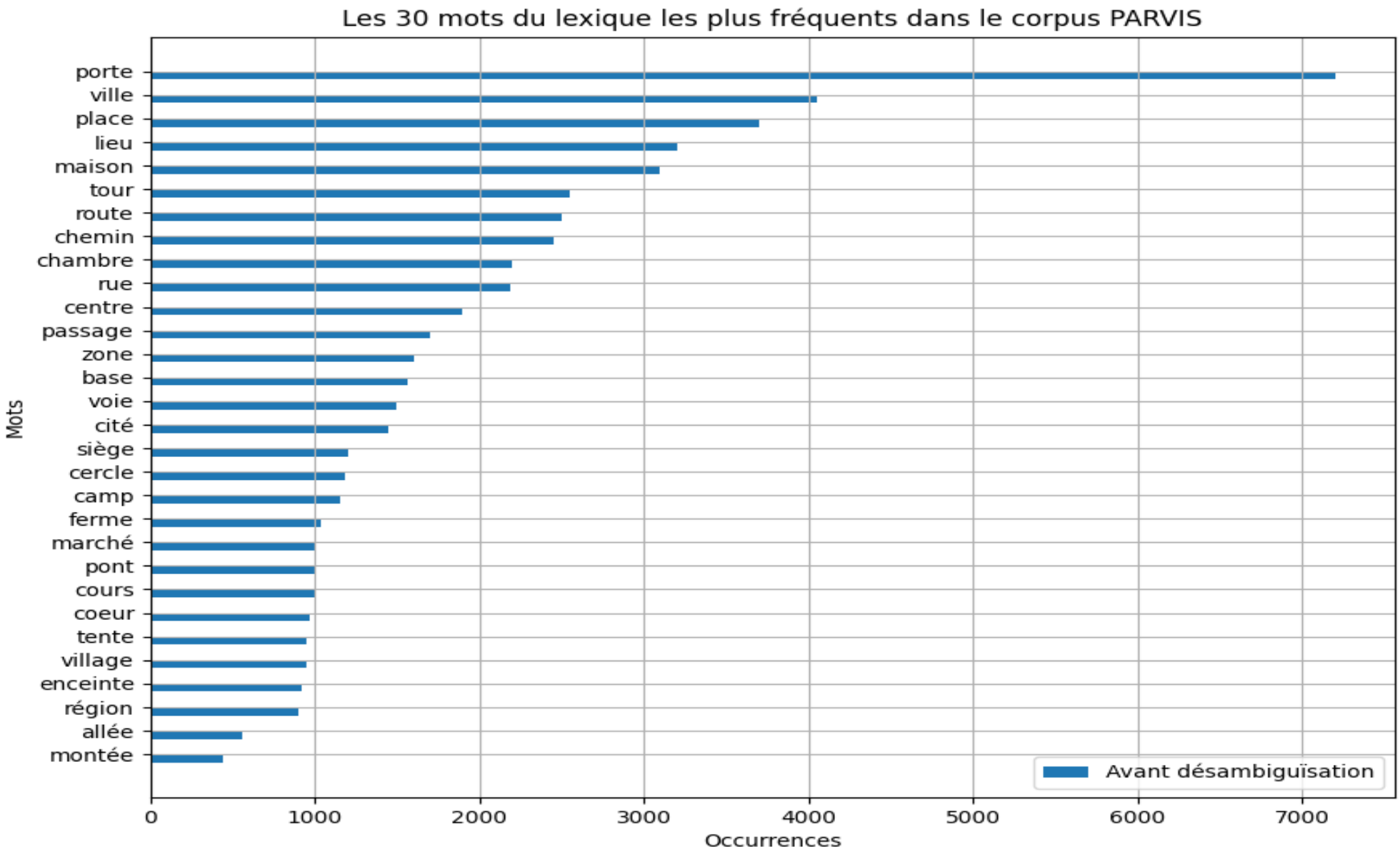
- La ressource terminologique a été construite comme un sous-ensemble du lexique de l'ouvrage « Les mots de la ville » [Topalov et al., 2010].

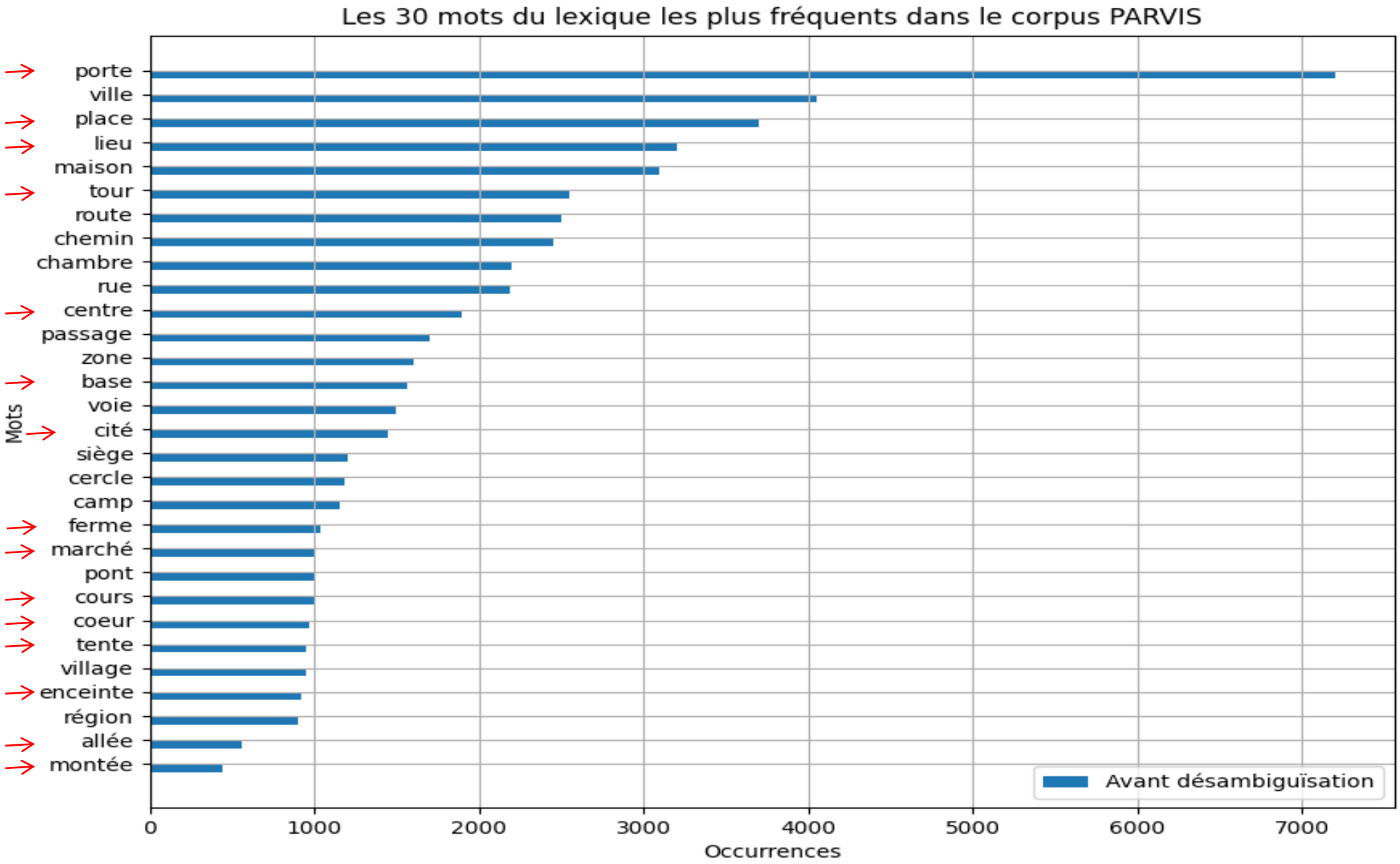
- L'ouvrage regroupe 533 mots, principalement des noms, désignant des éléments de la ville (lieux et objets).

- Seulement 183 mots du lexique sont employés dans le corpus PARVIS.

Problématique

- Certains mots de la ressource terminologique sont *polysémiques,* et ne sont pas toujours employés dans le contexte de la ville.


Les 30 mots du lexique les plus fréquents dans le corpus PARVIS
porte
ville
place
lieu
maison
tour
route
chemin
chambre
rue
centre
passage
zone
base
voie
cité
siège
cercle
camp
ferme
marché
pont
cours
coeur
tente
village
enceinte
région
allée
montée
Mots
0
1000
2000
3000
4000
5000
6000
7000
Occurrences
Avant désambiguïsation


Les 30 mots du lexique les plus fréquents dans le corpus PARVIS
porte
ville
place
lieu
maison
tour
route
chemin
chambre
rue
centre
passage
zone
base
voie
cité
siège
cercle
camp
ferme
marché
pont
cours
coeur
tente
village
enceinte
région
allée
montée
Mots
0
1000
2000
3000
4000
5000
6000
7000
Occurrences
Avant désambiguïsation

Problématique

- **Exemples :**

  - Le mot "cité" peut correspondre au nom "cité" (qui appartient au lexique de la ville), mais aussi être le participe passé du verbe "citer", qui n'appartient pas au lexique de la ville.

  - La polysémie est liée au contexte d'emploi du mot et non à son étiquette grammaticale. Par exemple, l'étiquette grammaticale du mot "centre" est la même dans les expressions «Le centre de la ville » et « Le centre de gravité» (dans les deux cas, c'est un nom), mais le mot n'est lié à la ville que dans la première expression.

Objectif

- Identifier les occurrences des mots de la ressource terminologique en lien avec le thème de la ville.

- Distinguer ceux employés dans le contexte de la ville de ceux qui ne le sont pas.

- Proposer une méthode de désambiguïsation des 15 mots polysémiques les plus fréquents.

Méthode

- La méthode proposée vise à donner, pour chaque occurrence de mot polysémique, une réponse booléenne à la question : "le contexte d'emploi de cette occurrence est-il associé à la ville ? ».

- Elle est fondée sur une méthode de classification.

- Pour chaque mot polysémique à désambiguïser, un classifieur est construit par apprentissage.

- Chaque classifieur classe une phrase contenant le mot polysémique sur lequel il est entraîné.

Méthode

- La construction des classifieurs repose sur :

1. L'annotation des phrases qui contiennent les mots polysémiques.
2. L'entraînement des classifieurs.

1- Annotation des phrases

Annotation des phrases

- Construction des jeux de données pour l'entraînement et le test des classifieurs.

- Annotation des phrases contenant les mots polysémiques pour la construction de ces jeux de données :

  - Attribution de l'étiquette "**1**" lorsque le mot polysémique de la phrase fait référence à la ville (comme : *Le virus se déplaçait d'un quartier de la **cité** à l'autre)* ,
  - Attribution de l'étiquette "**0**" dans le cas contraire (comme : *Elle l'a **cité** comme une sorte de réincarnation)* ,
  - Les phrases correspondant à des emplois figés ou métaphoriques du mot ont été annotées "**0**", (comme : *l'individualisme avait pris **place**).*

Jeux de données d'entraînement

- Constitution d'un jeu de données d'entraînement pour chacun des 15 mots polysémiques les plus fréquents dans le corpus PARVIS.
- Chaque jeu de données d'entraînement regroupe des phrases extraites du corpus PARVIS.
- Les phrases contiennent des exemples dans lesquels les mots polysémiques sont utilisés pour parler de la ville, ainsi que des exemples où ces mots ont d'autres sens.
- 80 % des phrases de ces jeux de données ont été destinées à l'entraînement des classifieurs (les 20 % restants constituent le jeu de test).

Jeux de données d'entraînement

Jeu de données PARVIS	# total de phrases	# de phrases avec l'étiquette "0"	# de phrases avec l'étiquette "1"
allée	80	41	39
base	63	35	28
centre	90	54	36
cité	80	27	53
cœur	50	30	20
cour	50	28	22
enceinte	40	18	22
ferme	40	21	19
lieu	56	28	28
marché	50	31	19
montée	60	39	21
place	50	28	22
porte	50	22	28
tente	50	22	28
tour	73	35	38

Table 1. *Description de chaque jeu de données d'entraînement*

Jeux de données de test

- **Objectif supplémentaire :** évaluer et éprouver la robustesse de la méthode.

- Jeux de données de test pour chacun des 15 mots polysémiques :
  - Des phrases extraites du corpus PARVIS.
  - Des phrases extraites du corpus du Grand Débat National (GDN).

- Caractéristiques du corpus du GDN :
  - Constitué de contributions écrites librement sur une plate-forme par des milliers de contributeurs et contributrices.
  - Moins de garanties en termes de syntaxe et de cohérence des phrases.

- Les phrases de test contiennent des exemples dans lesquels les mots polysémiques sont utilisés pour parler de la ville (étiquette “1”), ainsi que des exemples où ces mots ont d'autres contextes d’emploi (étiquette “0”).

2- Entraînement des classifieurs

Entraînement des classifieurs

- Chaque classifieur classe la phrase à l'aide des vecteurs de contexte (embeddings) de chacun de ses mots, calculés à l'aide du modèle de langue *CamemBERT* [Martin et al., 2019].

- Chaque classifieur est entraîné sur les phrases des jeux de données d'entraînement en affinant le modèle *CamemBERT* (fine-tuning) à l'aide de la fonction de classification par séquence du modèle (*Camembert For Sequence Classification*).

- L'architecture du modèle utilisée est *camembert-base (768 dimensions pour chaque vecteur).*

Évaluation des classifieurs

- Évaluation des classifieurs sur les jeux de données de test (les jeux de données du GDN, et 20% des jeux de données d'entraînement).
- Évaluation fondée sur deux indicateurs : l'**exactitude** (**accuracy**) et la **F- mesure**.
- Une moyenne d'accuracy de 96% sur les jeux de données de test de PARVIS.
- Une moyenne d'accuracy de 86% sur les jeux de données du GDN.

Évaluation des classifieurs

Corpus	PARVIS		GDN
Indicateur	**Accuracy**	**F-mesure**	**Accuracy**
allée	1.0	1.0	0.93
base	1.0	1.0	0.93
centre	0.85	0.83	1.0
cité	1.0	1.0	0.73
cœur	1.0	1.0	0.8
cour	0.9	0.93	0.93
enceinte	1.0	1.0	0.7
ferme	1.0	1.0	0.87
lieu	1.0	1.0	0.97
marché	1.0	1.0	0.8
montée	0.9	0.95	0.94
place	0.9	0.93	0.93
porte	0.9	0.93	0.73
tente	1.0	1.0	0.73
tour	1.0	1.0	0.87
Moyenne	**0.96**	**0.97**	**0.86**

Table 2. *Evaluation des classifieurs sur les jeux de données PARVIS et GDN*

Évaluation des classifieurs

Corpus	PARVIS		GDN
Indicateur	Accuracy	F-mesure	Accuracy
allée	1.0	1.0	0.93
base	1.0	1.0	0.93
centre	0.85	0.83	1.0
cité	1.0	1.0	0.73
cœur	1.0	1.0	0.8
cour	0.9	0.93	0.93
enceinte	1.0	1.0	0.7
ferme	1.0	1.0	0.87
lieu	1.0	1.0	0.97
marché	1.0	1.0	0.8
montée	0.9	0.95	0.94
place	0.9	0.93	0.93
porte	0.9	0.93	0.73
tente	1.0	1.0	0.73
tour	1.0	1.0	0.87
Moyenne	**0.96**	**0.97**	**0.86**

Table 2. *Evaluation des classifieurs sur les jeux de données PARVIS et GDN*

Évaluation des classifieurs

Corpus	PARVIS		GDN
Indicateur	**Accuracy**	**F-mesure**	**Accuracy**
allée	1.0	1.0	0.93
base	1.0	1.0	0.93
centre	0.85	0.83	1.0
cité	1.0	1.0	0.73
cœur	1.0	1.0	0.8
cour	0.9	0.93	0.93
enceinte	1.0	1.0	0.7
ferme	1.0	1.0	0.87
lieu	1.0	1.0	0.97
marché	1.0	1.0	0.8
montée	0.9	0.95	0.94
place	0.9	0.93	0.93
porte	0.9	0.93	0.73
tente	1.0	1.0	0.73
tour	1.0	1.0	0.87
Moyenne	**0.96**	**0.97**	**0.86**

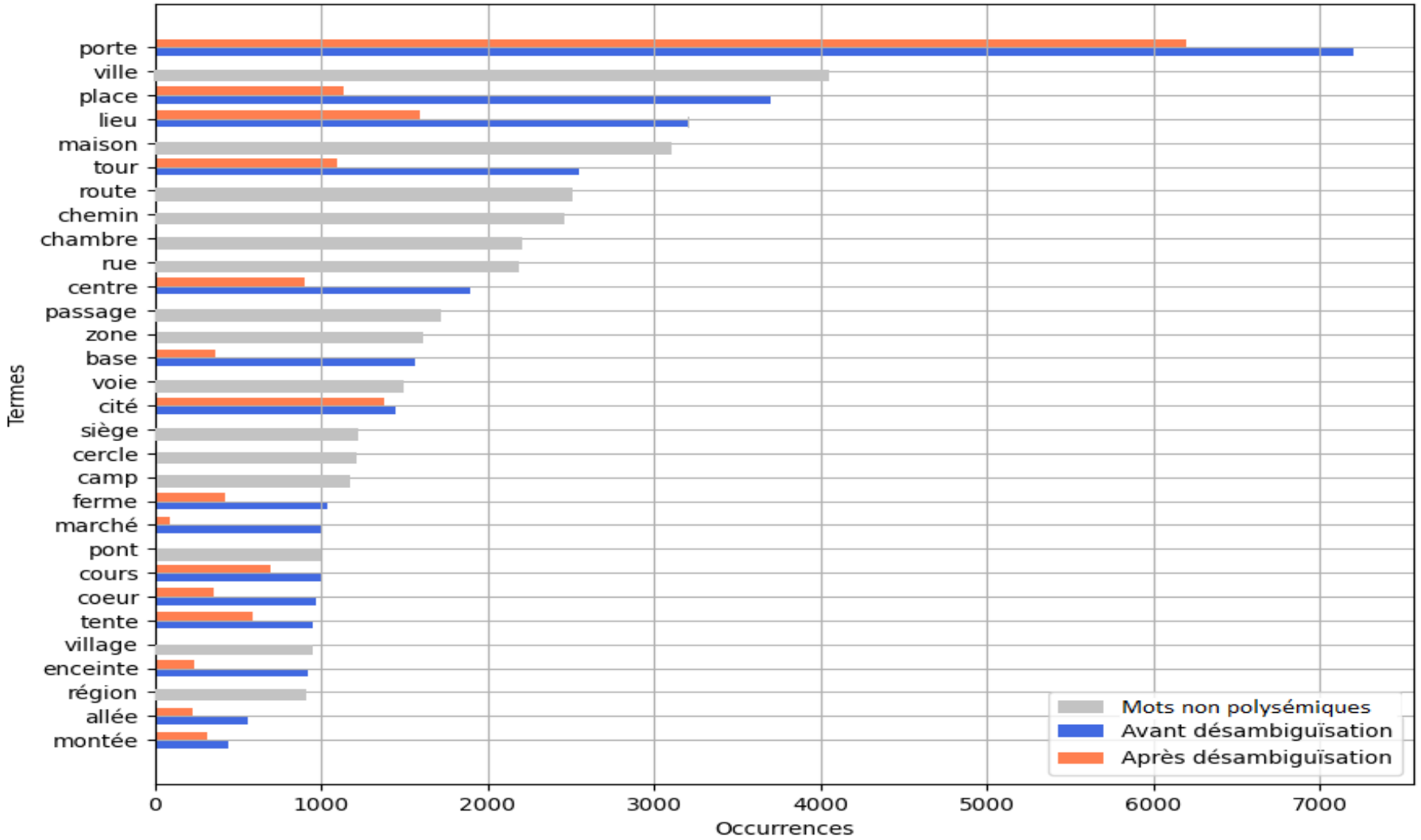
Table 2. *Evaluation des classifieurs sur les jeux de données PARVIS et GDN*

Évaluation des classifieurs

Corpus	PARVIS		GDN
Indicateur	**Accuracy**	**F-mesure**	**Accuracy**
allée	1.0	1.0	0.93
base	1.0	1.0	0.93
centre	0.85	0.83	1.0
cité	1.0	1.0	0.73
cœur	1.0	1.0	0.8
cour	0.9	0.93	0.93
enceinte	1.0	1.0	0.7
ferme	1.0	1.0	0.87
lieu	1.0	1.0	0.97
marché	1.0	1.0	0.8
montée	0.9	0.95	0.94
place	0.9	0.93	0.93
porte	0.9	0.93	0.73
tente	1.0	1.0	0.73
tour	1.0	1.0	0.87
Moyenne	**0.96**	**0.97**	**0.86**

Table 2. *Evaluation des classifieurs sur les jeux de données PARVIS et GDN*

Résultats de la désambiguïsation



Résultats de la désambiguïsation

Mot	# occurrences après désambiguïsation	rang avant désambiguïsation	rang après désambiguïsation
allée	226	29	62
base	364	14	48
centre	905	11	22
cité	1381	16	12
cœur	358	24	49
cour	698	23	27
enceinte	233	27	61
ferme	427	20	38
lieu	1596	4	11
marché	93	21	95
montée	317	30	50
place	1089	3	15
porte	6304	1	1
tente	591	25	31
tour	1077	6	16

Table 3. *Nombres d'occurrences des 15 mots polysémiques du lexique dans le corpus*

Résultats de la désambiguïsation

Mot	# occurrences après désambiguïsation	rang avant désambiguïsation	rang après désambiguïsation
allée	226	29	62
base	364	14	48
centre	905	11	22
cité	1381	16	12
cœur	358	24	49
cour	698	23	27
enceinte	233	27	61
ferme	427	20	38
lieu	1596	4	11
marché	93	21	95
montée	317	30	50
place	1089	3	15
porte	6304	1	1
tente	591	25	31
tour	1077	6	16

Table 3. *Nombres d'occurrences des 15 mots polysémiques du lexique dans le corpus*

Résultats de la désambiguïsation

Mot	# occurrences après désambiguïsation	rang avant désambiguïsation	rang après désambiguïsation
allée	226	29	62
base	364	14	48
centre	905	11	22
cité	1381	16	12
cœur	358	24	49
cour	698	23	27
enceinte	233	27	61
ferme	427	20	38
lieu	1596	4	11
marché	93	21	95
montée	317	30	50
place	1089	3	15
porte	6304	1	1
tente	591	25	31
tour	1077	6	16

Table 3. *Nombres d'occurrences des 15 mots polysémiques du lexique dans le corpus*

Résultats de la désambiguïsation

Mot	# occurrences après désambiguïsation	rang avant désambiguïsation	rang après désambiguïsation
allée	226	29	62
base	364	14	48
centre	905	11	22
cité	1381	16	12
cœur	358	24	49
cour	698	23	27
enceinte	233	27	61
ferme	427	20	38
lieu	1596	4	11
marché	93	21	95
montée	317	30	50
place	1089	3	15
porte	6304	1	1
tente	591	25	31
tour	1077	6	16

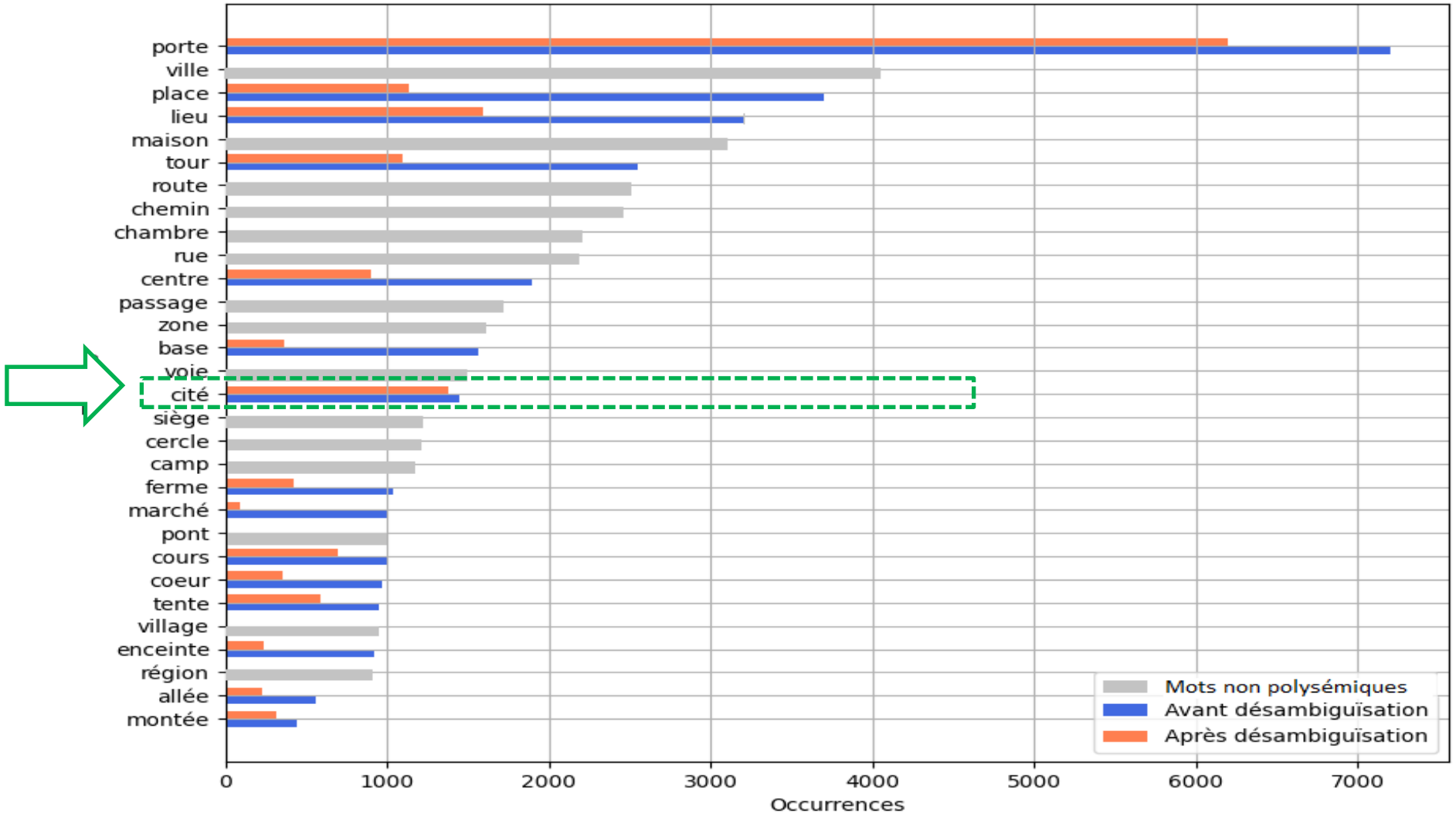
Table 3. *Nombres d'occurrences des 15 mots polysémiques du lexique dans le corpus*

Résultats de la désambiguïsation

Mot	# occurrences après désambiguïsation	rang avant désambiguïsation	rang après désambiguïsation
allée	226	29	62
base	364	14	48
centre	905	11	22
cité	1381	16	12
cœur	358	24	49
cour	698	23	27
enceinte	233	27	61
ferme	427	20	38
lieu	1596	4	11
marché	93	21	95
montée	317	30	50
place	1089	3	15
porte	6304	1	1
tente	591	25	31
tour	1077	6	16

Table 3. *Nombres d'occurrences des 15 mots polysémiques du lexique dans le corpus*

Résultats de la désambiguïsation



Résultats de la désambiguïsation

Mot	# occurrences après désambiguïsation	rang avant désambiguïsation	rang après désambiguïsation
allée	226	29	62
base	364	14	48
centre	905	11	22
cité	1381	16	12
cœur	358	24	49
cour	698	23	27
enceinte	233	27	61
ferme	427	20	38
lieu	1596	4	11
marché	93	21	95
montée	317	30	50
place	1089	3	15
porte	6304	1	1
tente	591	25	31
tour	1077	6	16

Table 3. *Nombres d'occurrences des 15 mots polysémiques du lexique dans le corpus*

Résultats de la désambiguïsation

Mot	# occurrences après désambiguïsation	rang avant désambiguïsation	rang après désambiguïsation
allée	226	29	62
base	364	14	48
centre	905	11	22
cité	1381	16	12
cœur	358	24	49
cour	698	23	27
enceinte	233	27	61
ferme	427	20	38
lieu	1596	4	11
marché	93	21	95
montée	317	30	50
place	1089	3	15
porte	6304	1	1
tente	591	25	31
tour	1077	6	16

Table 3. *Nombres d'occurrences des 15 mots polysémiques du lexique dans le corpus*

Résultats de la désambiguïsation

Mot	# occurrences après désambiguïsation	rang avant désambiguïsation	rang après désambiguïsation
allée	226	29	62
base	364	14	48
centre	905	11	22
cité	1381	16	12
cœur	358	24	49
cour	698	23	27
enceinte	233	27	61
ferme	427	20	38
lieu	1596	4	11
marché	93	21	95
montée	317	30	50
place	1089	3	15
porte	6304	1	1
tente	591	25	31
tour	1077	6	16

Table 3. *Nombres d'occurrences des 15 mots polysémiques du lexique dans le corpus*

Résultats de la désambiguïsation

Mot	# occurrences après désambiguïsation	rang avant désambiguïsation	rang après désambiguïsation
allée	226	29	62
base	364	14	48
centre	905	11	22
cité	1381	16	12
cœur	358	24	49
cour	698	23	27
enceinte	233	27	61
ferme	427	20	38
lieu	1596	4	11
marché	93	21	95
montée	317	30	50
place	1089	3	15
porte	6304	1	1
tente	591	25	31
tour	1077	6	16

Table 3. *Nombres d'occurrences des 15 mots polysémiques du lexique dans le corpus*

Résultats de la désambiguïsation

Mot	# occurrences après désambiguïsation	rang avant désambiguïsation	rang après désambiguïsation
allée	226	29	62
base	364	14	48
centre	905	11	22
cité	1381	16	12
cœur	358	24	49
cour	698	23	27
enceinte	233	27	61
ferme	427	20	38
lieu	1596	4	11
marché	93	21	95
montée	317	30	50
place	1089	3	15
porte	6304	1	1
tente	591	25	31
tour	1077	6	16

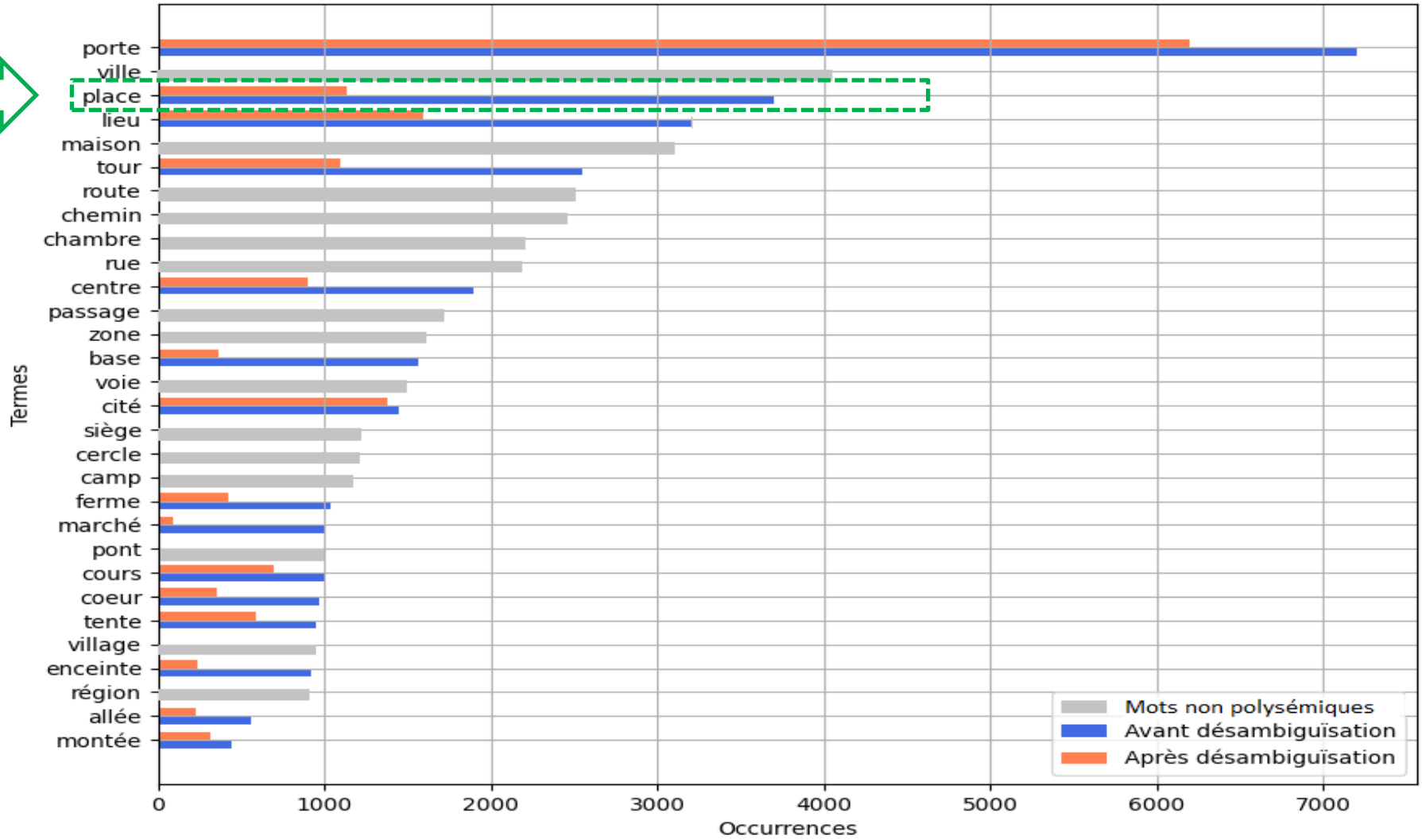
Table 3. *Nombres d'occurrences des 15 mots polysémiques du lexique dans le corpus*

Résultats de la désambiguïsation

Mot	# occurrences après désambiguïsation	rang avant désambiguïsation	rang après désambiguïsation
allée	226	29	62
base	364	14	48
centre	905	11	22
cité	1381	16	12
cœur	358	24	49
cour	698	23	27
enceinte	233	27	61
ferme	427	20	38
lieu	1596	4	11
marché	93	21	95
montée	317	30	50
place	1089	3	15
porte	6304	1	1
tente	591	25	31
tour	1077	6	16

Table 3. *Nombres d'occurrences des 15 mots polysémiques du lexique dans le corpus*

Résultats de la désambiguïsation



Conclusions et limites

- Le classifieur affichant la performance la moins élevée a été entraîné sur des phrases contenant le mot "**centre**" et a obtenu une exactitude de 85 %.

- Une explication serait que "**centre**" partage un trait sémantique de localisation avec d'autres contextes que celui de la ville, comme dans l'expression « *le centre de la terre* ».

- Limitations en termes de temps et de ressources, puisque, pour chaque nouveau mot à désambiguïser, une annotation de nouvelles phrases et un entraînement de nouveau classifieur sont nécessaires.

- L'évaluation a montré une robustesse de la méthode proposée, avec une exactitude de 86 % sur un corpus dont les variations lexicales, syntaxiques et stylistiques sont plus larges et moins prévisibles.

Merci de votre attention !

Question Time

Références

- Guembour S., Dong C., et Domingués C. (2023). Characterization of the city of the future from a science fiction corpus.
- Martin L., Muller B., Suárez P. J. O., Dupont Y., Romary L., Clergerie É. V., Seddah D., et Sagot B. (2019). Camembert : a tasty french language model.
- C. Topalov, L. C. de Lille, J.-C. Depaule, B. Marin (2010), L'aventure des mots de la ville.
- Bird S., Klein E., et Loper E. (2009). Natural language processing with Python : analyzing text with the natural language toolkit.