UNIVERSITÉ DE LORRAINE | BIBLIOTHÈQUES UNIVERSITAIRES

AVERTISSEMENT
DROIT D'AUTEUR

Ce document est le fruit d'un long travail approuvé par le jury de soutenance et mis à disposition de l'ensemble de la communauté universitaire élargie.

Il est soumis à la propriété intellectuelle de l'auteur. Ceci implique une obligation de citation et de référencement lors de l'utilisation de ce document. Toute contrefaçon, plagiat ou reproduction illicite encourt une poursuite pénale.

Contact bibliothèque : ddoc-theses-contact@univ-lorraine.fr
(Cette adresse ne permet pas de contacter les auteurs)

Code de la Propriété Intellectuelle. articles L 122. 4
Code de la Propriété Intellectuelle. articles L 335.2- L 335.10

UNIVERSITÉ DE LORRAINE
34 COURS LÉOPOLD — CS 25233
54052 NANCY CEDEX
TÉL : 03 54 50 54 00
CONTACT@UNIV-LORRAINE.FR
WWW.UNIV-LORRAINE.FR

UNIVERSITÉ DE LORRAINE

École doctorale SLTC Lorraine

Ressources lexicales et TAL : de l'enrichissement de réseaux lexico-sémantiques du français à la génération de plongements de graphe pour la désambiguïsation lexicale

THÈSE

présentée et soutenue publiquement le 20 avril 2026

pour l'obtention du

Doctorat de l'Université de Lorraine

(mention Sciences du Langage)

par

Hee-Soo CHOI

Composition du jury

Présidente :	Claire Gardent	LORIA, CNRS, Nancy, France
Rapporteur·ices :	Philippe Langlais	RALI, Université de Montréal, Canada
	Pascale Sébillot	IRISA, INSA Rennes, France
Examinateur·ices :	Lucie Barque	LLF, CNRS, Université Sorbonne Paris Nord, France
	Claire Gardent	LORIA, CNRS, Nancy, France
	Mathieu Lafourcade	LIRMM, Université Montpellier-II, Montpellier, France
Invité :	Bruno Guillaume	LORIA, Inria, Nancy, France
Encadrant·es:	Mathieu Constant	ATILF, CNRS, Université de Lorraine, Nancy, France
	Karën Fort	LORIA, CNRS, Université de Lorraine, Nancy, France

Analyse et Traitement Informatique de la Langue Française — UMR 7118

Mis en page avec la classe thesul.

Remerciements

J'ai eu la chance d'avoir été entourée d'un grand nombre de personnes qui ont marqué ces cinq ans de thèse et que je souhaite ici remercier. La liste est longue, mais cette thèse n'aurait jamais vu le jour sans vous.

Mes premiers remerciements sont destinés à mes trois encadrants : Mathieu Constant, Karën Fort et Bruno Guillaume. Merci Mathieu pour ta confiance et d'avoir pris le temps d'être présent malgré tes obligations de directeur de l'ATILF. Merci Bruno pour ta présence rassurante et d'avoir accepté de faire partie de mon encadrement même officieusement. Et enfin, merci Karën. Je n'aurais jamais imaginé, ni eu le courage de suivre ce chemin sans ce premier stage en M1, qui a tout changé, sur tous les plans. Merci pour ton investissement et ta générosité, ça a un impact bien plus important que tu ne l'imagines. Vous m'avez accompagnée durant ces longues années, avec patience et bienveillance. J'ai privilégié les personnes avec qui je voulais faire cette thèse et je ne le regrette pas. J'ai beaucoup appris à vos côtés. Merci.

Je tiens ensuite à remercier Philippe Langlais et Pascale Sébillot d'avoir accepté de rapporter cette thèse. Merci également à Lucie Barque, Mathieu Lafourcade et Claire Gardent d'avoir accepté de faire partie du jury. Je remercie particulièrement Claire Gardent et Christophe Benzitoun en tant que membres du comité de suivi.

Je remercie également les collègues de l'ATILF : Alain Polguère, Sandrine Ollinger, Veronika Lux-Pogodalla et Aman Sinha pour leur aide sur le RL-fr. Un grand merci à Hanbyul Song pour son amitié et nos conversations en Konglish.

Merci à tous les collègues de l'IDMC et de la FST avec qui j'ai pu travailler en tant qu'ATER. Un merci particulier à Damien Jamet et Jui-Ting Lu de m'avoir épargné des corrections en fin de rédaction de manuscrit et à Nasr Allah Aghelias, Colin Blake et Rémi Pallen pour leur compassion face à des salles de TP quasi vides.

Je souhaiterai ensuite remercier l'équipe Sémagramme que j'ai pu côtoyer pendant près de six ans. Merci aux membres permanents (non-précaires) : Philippe de Groote pour ta cuisine inégalable, Maxime Amblard pour nos conversations sur le métier de MCF, Vincent Martin de m'avoir rappelé toutes les semaines que j'avais une thèse à écrire et de m'avoir appris à présenter ma thèse à un chauffeur de bus, Sylvain Pogodalla et Jacques Jayez pour vos retours lors des répétitions de soutenance et Michel Musiol pour ton aide lors des auditions de bourse de thèse. J'ai une pensée particulière pour Guy Perrier, à qui je n'ai pas eu l'occasion d'expliquer ma thèse malgré son intérêt pour le sujet. Merci à tous les précaires de l'équipe : à la première génération, Maria, Lisa, Samuel, Amandine, Pierre. Un merci particulier à Pierre Lefebvre pour m'avoir poussé à « imposer mon style » dès le début de la thèse (ça aura fini par converger). À la deuxième génération avec qui j'ai pu partager tant de moments : Siyana, Valentin, Maxime, Vincent, Bertrand, Fanny, Amandine Bob, Marie, Clémentine, Rémi, Elio. Merci d'avoir été plus que des collègues et d'être devenus des amis.

Merci à toutes les personnes du Loria, que j'ai pu rencontrer à différentes occasions : à Cyril Regan, Hélène Gerecke et Nathan Gillot pour notre rivalité aux matchs de ping-pong. À Isabelle et Caro d'avoir rendu les pauses midi si agréables. À Marie Baron pour ton enthousiasme à tous nos projets bourbiers. À Georgios d'avoir partagé une partie de cette aventure entre Nancy,

Paris et Sitia. À Guilherme pour ta gentillesse et nos déjeuners chez les Michel. Aux meilleurs volleyeurs du SUAPS : Amandine, Julien, Diego et Sevra pour nos soirées au volley et au bar. À la team du tricot/crochet time : merci Marie de partager ta passion et tes talents (et pour tes délicieux canelés), merci à la meilleure professeure, Siyana de m'avoir fait tomber dans l'addiction. À mes demi-sœurs de thèse : Fanny, Clémentine et Fatou. À l'équipe Carbone d'avoir supporté mes passages dans leur bureau pour me plaindre, en particulier Victor Matrat, meilleur élève de crochet. Merci pour tes schémas et pour les check-ups de la météo de mes journées qui me redonnaient le sourire quand ça n'allait pas.

La thèse m'aura notamment permis de faire des belles rencontres en conférence. Merci à mes compatriotes de conférences de me faire rire à nos rendez-vous annuels à TALN : Dr. Paul Lerner et tes questions à toutes les présentations, Dr. Antoine Chaffin, meilleur partenaire de pétanque, Dr. Shu Okabe, best drinking buddy de conférence.

I also want to thank the SapienzaNLP team in Rome, with whom I spent three incredible months. Thank you for welcoming me so warmly and for becoming true friends. 2024 has deeply changed me, and it is largely thanks to you all. Grazie mille, ragazzi.

Je remercie également mes amis parisiens qui ont suivi cette aventure de loin. Mes précieuses amies de l'époque de médecine : Odile pour nous faire toujours autant rire, Xuan pour ton grand cœur et ta joie de vivre et Amélie pour ta douceur et ton écoute. Merci aux amis du lycée qui sont encore là après tant d'années : Léo, Arthur, Clément, Hector et Antonin. Un merci particulier à Charly pour ton amitié après toutes ses années et pour tous les restos payés parce que je n'étais qu'une étudiante précaire.

Je tiens aussi à remercier la famille Michel, pour nos déjeuners réguliers. Merci d'avoir été une figure de grands-parents pour moi en France et merci Véronique d'avoir été présente à ma soutenance.

Et plus particulièrement, merci à la LREC team : Fanny pour ta gentillesse et ton empathie sans égale, Siyana pour tes "you'll be fine" réconfortants, Priyansh d'avoir sauvé ma thèse avec cette collaboration légendaire. Merci à mes traquenardeurs préférés, Bebert et Maxou, pour tous les moments à l'escalade, au footing du dimanche matin, autour d'une table de ping-pong ou d'une (ou plusieurs...) bouteille(s) de porto.... Merci à ma team Rocket, Vincent et Gabriel pour avoir été mon plus gros soutien. Merci Vincent d'être toujours prêt à aider les autres, souvent à tes dépens, d'avoir réglé toutes mes galères techniques, pour toutes nos parties de ping-pong, tous nos gâteaux d'anniversaire et le 100% sur les stations de métro. Je te dois beaucoup et j'espère un jour pouvoir te rendre la pareille.

Merci Gabriel, d'avoir été présent durant les bons moments, mais surtout durant les plus difficiles. Merci d'avoir toujours eu les mots justes et d'avoir cru en moi lorsque je n'y arrivais plus moi-même. Je ne serais jamais arrivée au bout de cette thèse sans ton soutien et ton aide inconditionnels. Merci du fond du cœur.

Enfin je terminerai par remercier ma famille. Je ne saurais sans doute jamais l'ampleur des sacrifices que vous avez faits en restant en France, mais je n'oublierai pas. Merci pour tout, cette réussite est tout autant la vôtre.

Table des matières

Table des figures **ix**

Liste des tableaux **xi**

Résumé **1**

Abstract **3**

Publications **5**

Introduction **7**

1 Enjeu pour le TAL : comment et pourquoi représenter les connaissances linguistiques ? 8

2 Enjeu pour les ressources lexicales : comment valoriser leur caractère explicite ? 9

3 Cadre de la thèse 10

Partie I Les ressources lexicales en Traitement Automatique des Langues 13

Introduction **15**

Chapitre 1 De la linguistique au traitement automatique des langues **17**

1.1 Introduction 17

1.2 Qu'est-ce qu'une ressource lexicale ? 18

1.2.1 L'ambiguïté du mot 19

1.2.2 Type de ressource et contenu linguistique 19

1.2.3 Caractéristiques d'une ressource lexicale 21

1.3 Les ressources lexicales : empreintes des théories linguistiques 22

1.3.1 Le Lexique-Grammaire et le distributionnalisme 22

1.3.2 Dicovalence : notion de valence et approche pronominale 23

1.3.3 Le Réseau lexical du français et la théorie Sens-Texte 24
1.3.4 La famille des *Net* : WordNet, FrameNet, VerbNet 25
1.3.5 L'influence de la linguistique générative 26
1.4 Les ressources lexicales : « matériau » pour le TAL 27
1.4.1 Briques fondamentales des approches symboliques et statistiques . . . 27
1.4.2 Construction automatique : compromis entre couverture et qualité . . 28
1.4.3 Pérennité et accessibilité des ressources lexicales 31
1.5 Les plongements lexicaux : une représentation neuronale des connaissances linguistiques 32
1.5.1 Plongements lexicaux statiques 32
1.5.2 Plongements lexicaux contextuels 33
1.6 La place des ressources lexicales aujourd'hui en TAL 33
1.6.1 Intégration de ressources lexicales aux grands modèles de langue . . . 33
1.6.2 Enjeux pour les langues peu dotées 35
1.6.3 Les grands modèles de langue ont-ils des « compétences lexicales » ? . 35
1.7 Conclusion 36

Chapitre 2 25 ans de ressources lexicales du français à TALN 37
2.1 Introduction 37
2.2 Des ressources lexicales du français libres d'utilisation 38
2.2.1 Ressources lexicales classiques 38
2.2.2 Plongements lexicaux statiques 39
2.3 Le corpus TALN 2001-2025 40
2.3.1 Constitution du corpus 40
2.3.2 Pré-traitements et annotations manuelles 41
2.4 Évolution diachronique des ressources 42
2.5 Évolution des applications 45
2.5.1 Construction de ressources spécialisées 45
2.5.2 Construction d'outils 45
2.5.3 Correction d'erreurs et évaluation 46
2.5.4 Des ressources symboliques aux plongements lexicaux statiques 47
2.6 La question de la réutilisabilité 47
2.7 Conclusion 49

Partie II Enrichir des réseaux lexico-sémantiques par prédiction de liens 51

Introduction **53**

Chapitre 3 État de l'art : complétion de graphes par prédiction de liens **55**

3.1 Introduction . . . 55
3.2 Graphes de connaissances et réseaux lexico-sémantiques . . . 56
3.2.1 Définitions et notations . . . 56
3.2.2 Contexte, construction et applications . . . 58
3.3 Pallier l'incomplétude des graphes de connaissances . . . 61
3.3.1 Complétion externe . . . 61
3.3.2 Complétion interne . . . 62
3.4 Prédiction de liens . . . 63
3.4.1 Description de la tâche . . . 63
3.4.2 Jeux de données . . . 64
3.4.3 Métriques d'évaluation . . . 65
3.4.4 Méthodes de prédiction de liens . . . 66
3.5 Conclusion . . . 70

Chapitre 4 Génération de prédictions robustes pour l'enrichissement de réseaux lexico-sémantiques **71**

4.1 Enjeux et contributions . . . 71
4.2 Réseaux lexico-sémantiques du français . . . 73
4.2.1 De `RezoJDM` à `RezoJDM16k` . . . 73
4.2.2 Le `RL-fr`, Réseau Lexical du Français . . . 74
4.2.3 Deux réseaux lexico-sémantiques différents mais complémentaires . . . 75
4.3 Prédiction de liens . . . 76
4.3.1 Pré-traitement des jeux de données . . . 76
4.3.2 Modèles et métriques . . . 77
4.3.3 Comparaison des résultats . . . 78
4.4 Robustesse du modèle à travers l'incertitude prédictive . . . 80
4.4.1 Conception et entraînement d'un réseau de neurones profond . . . 80
4.4.2 Modèle formel de l'incertitude prédictive . . . 81
4.4.3 Approximation de l'incertitude avec le Monte-Carlo Dropout . . . 82
4.5 Enrichir un réseau lexico-sémantique parcimonieux . . . 84
4.5.1 Application du Monte-Carlo Dropout . . . 84
4.5.2 Sélection de triplets candidats . . . 84
4.5.3 Annotations manuelles . . . 85

4.5.4 Extraction de nouveaux triplets 87
4.6 Conclusion 88

Partie III Évaluer l'intégration de plongements de graphe pour la désambiguïsation lexicale 91

Introduction 93

Chapitre 5 Désambiguïsation lexicale : état de l'art et biais d'évaluation 95
5.1 Contexte et description 95
5.2 De Senseval à SemEval : campagnes d'évaluation pour la désambiguisation lexicale 97
5.2.1 Senseval 97
5.2.2 SemEval 98
5.3 Données annotées 99
5.3.1 Annotations en sens WordNet/BabelNet 100
5.3.2 Annotations avec d'autres inventaires de sens 103
5.4 Approches principales 104
5.4.1 Approches basées sur les connaissances 104
5.4.2 Approches neuronales supervisées 105
5.4.3 Approches avec les grands modèles de langue auto-régressifs 108
5.5 Métriques traditionnelles 110
5.6 Une évaluation biaisée par un cadre avantageux 111
5.6.1 Biais de la composition des données d'évaluation 112
5.6.2 Biais de la métrique 115
5.6.3 Biais du sens le plus fréquent et des données d'entraînement 116
5.6.4 Des performances dépendantes des données 117
5.7 Conclusion 120
5.7.1 Les limites de la tâche de classification 120
5.7.2 Un cadre d'évaluation « trop » ancré 121

Chapitre 6 Intégration de plongements de graphe et évaluation de la désambiguïsation lexicale 123
6.1 Enjeux et contributions 124
6.2 `BEL-RL-fr`, un corpus atypique pour la désambiguïsation lexicale 125
6.2.1 Exemples lexicographiques annotés avec le `RL-fr` 125
6.2.2 Pré-traitement des données 126

6.3 Systèmes de désambiguïsation lexicale 128
6.3.1 *Baselines* 128
6.3.2 Modèle `EWISER` et intégration de plongements de graphe 131
6.3.3 Grands modèles de langue 133
6.4 Évaluer au-delà de la classification 136
6.4.1 Métriques 136
6.4.2 Évaluation restreinte *vs* Évaluation globale 136
6.5 QR 1 : performances des systèmes en évaluation restreinte 138
6.5.1 Meilleures performances avec un modèle neuronal supervisé 139
6.5.2 Des données d'évaluation qui influencent les métriques 140
6.5.3 Conclusion de la QR 1 141
6.6 QR 2 : influence de l'intégration de connaissances 141
6.6.1 Comparaison des prédictions des modèles 141
6.6.2 Une baisse des scores en évaluation globale 144
6.6.3 Une distribution de probabilités déviée par l'intégration de connaissances 147
6.6.4 Des performances variables selon le niveau de polysémie 149
6.6.5 Conclusion de la QR 2 152
6.7 QR 3 : prédiction des sens non vus à l'entraînement 152
6.7.1 Corrélation entre la distribution des sens à l'entraînement et la performance des modèles 153
6.7.2 Amélioration des performances des sens non vus avec `EWISER+A` 153
6.7.3 Conclusion de la QR 3 154
6.8 QR 4 : analyse des performances des grands modèles de langue 155
6.9 Discussion et conclusion 157

Conclusion **161**

Conclusion et perspectives **163**
1 Enrichissement mutuel entre ressources lexicales et systèmes de TAL 163
2 Nécessité d'évaluer au-delà d'une métrique unique 165
3 Changement de statut des ressources lexicales en TAL 166
4 Perspectives 167

Bibliographie **169**

Bibliographie **171**

Table des figures

1.1 Extrait de la table 33 des verbes (Figure 1 de Tolone (2009)). 23
1.2 Exemple du verbe « donner » dans Dicovalence. 24
1.3 Schéma du modèle sens-texte et des structures composant les représentations sémantiques et syntaxiques profondes (Figure 1 de Cousin (2023)). 24
1.4 Exemple de relation entre *synsets* 25
1.5 Ressources lexicales du français selon la méthode de construction. 28

2.1 Nombre d'articles du corpus TALN 2001-2025 par année. 41
2.2 Distribution dans le temps du nombre d'articles présentant une ressource lexicale traditionnelle selon le type d'usage. 43
2.3 Distribution dans le temps du nombre d'articles présentant des plongements lexicaux selon le type d'usage. 43
2.4 Distribution des ressources symboliques les plus fréquentes selon le type d'usage. 44
2.5 Distribution des plongements lexicaux selon le type d'usage. 44

3.1 Exemple d'un graphe de connaissance (gauche) et d'un réseau lexico-sémantique (droite). 57
3.2 Aperçu des graphes de connaissances populaires. Le tableau présente le nombre d'entités (*instances*), de triplets (*facts*), de types d'entités (*Types*) et de types de relations (*Relations*) (Tableau 1 de Paulheim (2017)). 59
3.3 Taxonomie des tâches impliquant les graphes de connaissances dans la littérature en TAL (Figure 1 de Schneider et al. (2022)). 60
3.4 Schéma des phases d'entraînement et d'inférence de la prédiction de liens. 64
3.5 Méthodes de complétion de graphe. 66

4.1 Exemple d'une partie de `JeuxDeMots` où le participant doit entrer des termes associés au terme `consentement`. Le dernier terme entré est `accord` et le jeu présente ses différents sens (raffinements). Le participant peut raffiner sa réponse en choisissant un des sens proposés. 73
4.2 Représentation de la copolysémie dans le `RL-fr`. 75
4.3 Vocable `accord` représenté dans le `RL-fr`. 76
4.4 Prédictions pour la tête `bonnet I` et la relation `synonymie`. Les triplets existant déjà dans le `RLF27k` sont précisés avec les couleurs : en bleu, ceux dans l'échantillon d'entraînement ([tr] pour *training*) et en rouge, ceux dans l'échantillon de test ([ts] pour test). L'axe des ordonnées est mis à l'échelle logarithme pour une meilleure lisibilité. 79
4.5 Exemple d'un type de triplet candidat dans `RLF27k`. Les flèches pleines représentent les relations existantes et celles en pointillé, les relations potentielles. . . . 85

4.6 Exemple d'un type de triplet candidat dans `RezoJDM16k`. Les flèches pleines représentent les relations existantes et celle en pointillé, une relation potentielle. . . 85
4.7 Kappa de Cohen sur `RLF27k` et `RezoJDM16k`. 86
4.8 Corrélations entre les valeurs d'annotation et les scores de confiance dans `RLF27k` et de `RezoJDM16k`. Les triplets considérés sont ceux pour lesquels les annotateurs sont d'accord. 87
4.9 Ratio de prédictions correctes (annotées 1) du `RLF27k` selon le seuil de confiance. 88

5.1 Annotations d'une phrase en sens `WordNet` dans ALL_{EN} (senseval2.d000.s008). . 99
5.2 Statistiques des jeux de données de désambiguïsation lexicale des campagnes Senseval et SemEval après unification (Table 1 de Raganato et al. (2017)). 101
5.3 *Pipeline* classique d'un modèle neural supervisé de désambiguïsation lexicale. . . 105
5.4 Méthode des k plus proches voisins pour la désambiguïsation lexicale. Les plongements de sens et les plongements contextuels de la phrase d'entrée sont dans le même espace vectoriel. (Figure 1 de Loureiro and Jorge (2019)). 106
5.5 Schéma du modèle `ESCHER` pour l'extraction de séquence (diapositive 10 de la présentation (Barba et al., 2021a)). 107
5.6 Distribution du nombre d'instances par sens dans ALL_{EN} et `42D`. 119

6.1 Annotation d'un exemple lexicographique dans le `BEL-RL-fr` v2.0 (exemple 129). 125
6.2 Distribution du nombre de lexèmes en fonction du nombre de sens dans le `BEL-RL-fr`. 127
6.3 Distribution des sens selon leur fréquence dans l'échantillon d'entraînement pour les noms (à gauche) et pour les verbes (à droite). 129
6.4 *Prompt* en français pour la désambiguïsation lexicale. Les sens sont décrits avec le voisinage lexical. 135
6.5 Distribution du nombre d'instances par sens dans l'échantillon de test pour les noms et les verbes. 140
6.6 Nombres d'instances correctement prédites en évaluation restreinte ($gold = \hat{s}_{\mathcal{S}(w)}$) selon les modèles `EWISER` sur les noms. 146
6.7 Nombres d'instances correctement prédites en évaluation restreinte (gold $= \hat{s}_{\mathcal{S}(w)}$) selon les modèles `EWISER` sur les verbes. 146
6.8 Top 10 des sens prédits par `EWISER-A` sur l'instance cit15678. Le sens *gold* est indiqué en rouge. 149
6.9 Top 10 des sens prédits par `EWISER+A COS` sur l'instance cit15678. Le sens *gold* est indiqué en rouge. 150
6.10 Performances des modèles en micro-F1 en évaluation restreinte sur les noms selon le niveau de polysémie. Le nombre d'instances est indiqué entre parenthèses. . . . 151
6.11 Performances des modèles en micro-F1 en évaluation restreinte sur les verbes selon le niveau de polysémie. Le nombre d'instances est indiqué entre parenthèses. . . . 151

Liste des tableaux

1.1 Ressources mentionnées dans le chapitre 1 et librement disponibles. 32

2.1 Liste des 26 ressources lexicales traditionnelles regroupées par domaine et par type. 39
2.2 Créateurs des ressources les plus utilisées dans le corpus TALN 2001-2025. 48
2.3 Nombre d'utilisations internes ou externes pour les quatre ressources les plus utilisées dans le corpus TALN 2001-2025. 49

3.1 Aperçu de la composition des réseaux lexico-sémantiques populaires. 60
3.2 Jeux de données pour la prédiction de liens. 64

4.1 Statistiques sur `RezoJDM16k` et `RL-fr`. 76
4.2 Statistiques sur les échantillons d'entraînement, validation et test de `RezoJDM16k` et `RLF27k` après une division en 80 %, 10 %, 10 %. 77
4.3 Description des modèles de prédiction de liens utilisés. $\Re$ est l'opérateur « partie réelle ». 78
4.4 Résultats des modèles de prédiction de liens sur `RezoJDM16k`. 78
4.5 Résultats des modèles de prédiction de liens sur `RLF27k`. 79
4.6 Exemples de triplets candidats de `RLF27k` composés d'une autre acception d'un triplet existant. 88
4.7 Exemples de triplets candidats de `RLF27k` composés d'une autre acception d'un triplet existant. 88
4.8 Exemples de triplets de `RezoJDM16k` avec une relation affinée. 89

5.1 Tableau récapitulatif des données principales disponibles pour la désambiguïsation lexicale. 100
5.2 Exemples de *prompts* testés par Meconi et al. (2025). 109
5.3 Exemple de classification à classes multiples avec l'approche « un contre le reste », « reste » signifiant n'importe quelle autre classe. 111
5.4 Exemple d'un calcul d'une micro-F1 dans une classification à classes multiples. . 111
5.5 Performances (%) en micro-F1 des principaux modèles de désambiguisation lexicale. 112
5.6 Performances en micro-F1 sur ALL_{EN} et $ALL_{ENnoMono}$. Nous avons ajouté les résultats de la ligne en gris. 113
5.7 Performances (%) en micro-F1 sur ALL_{EN}* et ALL_{NEW}. ALL_{EN}* correspond au sous-ensemble de ALL_{EN} dont les instances sont également dans ALL_{NEW} (Tableau 5 de Maru et al. (2022)). *#multi-gold* correspond au nombre d'instances présentant plusieurs sens corrects. 114
5.8 Statistiques sur les instances *multi-gold* dans ALL_{EN}* et ALL_{NEW}. 114
5.9 Exemple de résultats de classification avec des classes non équilibrées. 115

5.10 Performances (%) en macro-F1 (M-F1) et micro-F1 (m-F1), avec deltas (Δ = m-F1 - M-F1), sur ALL_{EN}, $ALL_{ENnoMono}$ et ALL_{NEW} (Tableau 1 de Maru et al. (2022)). 116
5.11 Pourcentage de fois que les systèmes prédisent le sens le plus fréquent (Tableau 1 de Maru et al. (2022)). 117
5.12 Pourcentage de fois que les systèmes prédisent un sens qui apparaît au moins une fois dans l'échantillon d'entraînement (Tableau 1 de Maru et al. (2022)). 117
5.13 Performances (%) en macro-F1 (M-F1) et micro-F1 (m-F1), avec deltas (Δ = m-F1 - M-F1) des systèmes sur ALL_{EN}, $ALL_{ENnoMono}$, ALL_{ENnoSC}, $ALL_{ENnoSCnoMono}$, $ALL_{ENno1st}$ et 42D (Tableau 4 de Maru et al. (2022)). 118
5.14 Statistiques sur la distribution des sens dans ALL_{EN} et `42D`. 119

6.1 Représentation de la polysémie du lexème `avocat` dans le `RL-fr`. 126
6.2 Statistiques sur les exemples lexicographiques du `BEL-RL-fr` (v2). 126
6.3 Nombre d'instances dans les échantillons d'entraînement, de validation et de test après division. 128
6.4 Exemple de la *baseline* basée sur le voisinage lexical. 130
6.5 Configurations du modèle `EWISER` selon les informations relatives au graphe. . . . 133
6.6 Deux types de description des sens possibles du lexème `bicyclette`. 134
6.7 Stratégies de *prompting* adoptées dans Meconi et al. (2025). 135
6.8 Performance (%) en macro-F1 (M-F1) et micro-F1 (m-F1) des modèles de désambiguïsation lexicale sur `BEL-RL-fr` sur les noms et les verbes en évaluation restreinte. Les meilleurs résultats pour chaque configuration sont présentés en gras. 138
6.9 Stratégie de *prompting* donnant les meilleurs résultats en micro-F1 selon les modèles. 139
6.10 Top 3 des sens s_i et des probabilités $P(s_i)$ des modèles `EWISER` sur l'instance cit6158 ayant pour *gold* = `chausson I.1`. 142
6.11 Top 3 des sens s et des probabilités des modèles `EWISER` sur l'instance cit19262 ayant pour *gold* = `laver II`. 143
6.12 Cas de figure des prédictions selon l'évaluation globale ou l'évaluation restreinte. 144
6.13 Performances (%) en macro-F1 (M-F1) et micro-F1 (m-F1) du modèle `EWISER` avec et sans intégration de connaissances relatives au graphe sur les noms. Les meilleurs résultats sont donnés en gras. 145
6.14 Performances (%) en macro-F1 (M-F1) et micro-F1 (m-F1) du modèle `EWISER` avec et sans intégration de connaissances relatives au graphe sur les verbes. Les meilleurs résultats sont donnés en gras. 145
6.15 Prédictions et probabilités des modèles `EWISER` sur l'instance cit6208 ayant pour *gold* = `avocat 1.II`. Dist($\hat{s}_{\mathcal{S}(w)}$, $\hat{s}_{\mathcal{V}}$) représente la distance entre $\hat{s}_{\mathcal{S}(w)}$ et $\hat{s}_{\mathcal{V}}$ dans le `RL-fr` en nombre de pas. 148
6.16 Statistiques sur les instances prédites correctement avec la probabilité maximale par `EWISER-A` et `EWISER+A COS`. 148
6.17 Statistiques sur les instances de test dont le sens *gold* est un sens vu ou non vu à l'entraînement. 153
6.18 Performances (%) en micro-F1 des modèles `EWISER` sur les instances des noms dont le sens correct est vu ou non vu à l'entraînement. 154
6.19 Performances (%) en micro-F1 des modèles `EWISER` sur les instances des verbes dont le sens correct est vu ou non vu de l'entraînement. 155
6.20 Performances (%) en macro-F1 (M-F1) et micro-F1 (m-F1) des grands modèles de langue sur les noms avec les stratégies **lexical** et **lexical_def**. 156

6.21 Performances (%) en macro-F1 (M-F1) et micro-F1 (m-F1) des grands modèles de langue sur les noms selon les stratégies **usage** et **usage_def**. 156
6.22 Performances (%) en macro-F1 (M-F1) et micro-F1 (m-F1) des grands modèles de langue sur les verbes selon les stratégies **lexical** et **lexical_def**. 156

Résumé

Cette thèse explore la relation bilatérale qui unit les ressources lexicales et le Traitement Automatique des Langues (TAL). Historiquement, les ressources lexicales symboliques, telles que les dictionnaires et les réseaux lexico-sémantiques, constituaient les fondations des systèmes de TAL. Cependant, l'essor de l'apprentissage profond et des grands modèles de langue a provoqué un changement de paradigme majeur vers des représentations implicites et numériques, apprises à partir de grandes quantités de données. Cette évolution soulève des défis cruciaux : un manque de transparence et d'interprétabilité des modèles, une représentation souvent déficiente des phénomènes linguistiques rares, et une aggravation des inégalités entre les langues bien dotées et celles peu dotées. Cette thèse s'articule autour de deux axes majeurs : explorer comment les techniques de TAL peuvent contribuer à automatiser et enrichir les ressources lexicales, et inversement, examiner comment l'intégration de ces ressources peut améliorer les systèmes automatiques.

La première partie de la thèse présente un panorama historique de l'évolution des ressources lexicales. Le chapitre 1 retrace la trajectoire de ces ressources, de leurs origines théoriques en linguistique jusqu'à leur application pratique dans le TAL moderne. Il met en lumière la transition d'un format symbolique lisible par l'humain vers un format numérique conçu pour la machine, modifiant intrinsèquement la manière dont l'information linguistique est représentée. Le chapitre 2 complète ce cadre théorique par une analyse bibliométrique diachronique des ressources lexicales du français sur une période de 25 ans, basée sur les actes de la conférence nationale TALN. Cette étude observe le cycle de vie de diverses ressources, retraçant l'évolution de leur utilisation, la transition vers les représentations vectorielles, et les enjeux persistants de réutilisabilité, d'accessibilité et de pérennité des ressources lexicales.

La deuxième partie de la thèse se concentre sur la dimension « TAL pour les ressources », en abordant la problématique de l'incomplétude inhérente aux graphes de connaissances et aux réseaux lexico-sémantiques. Après avoir dressé un état de l'art sur la complétion de graphes et la prédiction de liens au chapitre 3, le chapitre 4 applique ces techniques à deux réseaux du français, le `RL-fr` (Réseau Lexical du Français), ressource de granularité fine construite manuellement, et `RezoJDM`, une ressource très dense obtenue par production participative. Plutôt que de chercher à maximiser les performances, nous adoptons une démarche orientée vers la ressource. En exploitant la technique du Monte-Carlo Dropout pour estimer l'incertitude prédictive d'un modèle de prédiction de liens, nous générons des triplets candidats robustes pour une potentielle intégration dans la ressource.

La troisième partie explore l'axe inverse, « Ressources pour le TAL », en cherchant à déterminer si l'intégration de connaissances structurelles issues d'un réseau lexico-sémantique peut améliorer la désambiguïsation lexicale. Le chapitre 5 présente un état de l'art de la tâche de désambiguïsation lexicale et une analyse de son cadre d'évaluation traditionnel. Dans le chapitre 6, nous évaluons différents systèmes de désambiguïsation lexicale dont l'intégration de plongements de graphe, appris à partir du `RL-fr`, dans un modèle neuronal supervisé.

Dans cette thèse, nous nous sommes intéressés à l'enrichissement mutuel entre les ressources lexicales et les systèmes de TAL. Bien que les techniques contemporaines de TAL ont évolué vers l'utilisation de grands modèles de langue appris sur de grandes quantités de textes, nous insistons sur l'utilité des ressources lexicales, par leur description contrôlée et précise des phénomènes linguistiques.

Abstract

This thesis explores the bilateral relationship between lexical resources and Natural Language Processing (NLP). Historically, symbolic lexical resources, such as dictionaries and lexical-semantic networks, formed the foundation of NLP systems. However, the rise of deep learning and large language models has caused a major paradigm shift towards implicit, numerical representations learned from large amounts of data. This evolution raises crucial challenges : a lack of transparency and interpretability of models, an often deficient representation of rare linguistic phenomena, and a widening gap between well-endowed and poorly endowed languages. This thesis focuses on two main areas : exploring how NLP techniques can contribute to automating and enriching lexical resources, and conversely, examining how the integration of these resources can improve automatic systems.

The first part of the thesis presents a historical overview of the evolution of lexical resources. Chapter 1 traces the trajectory of these resources, from their theoretical origins in linguistics to their practical application in modern NLP. It highlights the transition from a human-readable symbolic format to a machine-designed digital format, intrinsically changing the way linguistic information is represented. Chapter 2 complements this theoretical framework with a diachronic bibliometric analysis of French lexical resources over a 25-year period, based on the proceedings of the TALN national conference. This study observes the life cycle of various resources, tracing the evolution of their use, the transition to vector representations, and the persistent challenges of reusability, accessibility, and sustainability of lexical resources.

The second part of the thesis focuses on the "NLP for resources" dimension, addressing the issue of incompleteness inherent in knowledge graphs and lexical-semantic networks. After reviewing the state of the art in graph completion and link prediction in Chapter 3, Chapter 4 applies these techniques to two French networks : `RL-fr` (Réseau Lexical du Français), a manually constructed fine-grained resource, and `RezoJDM`, a very dense resource obtained through crowdsourcing. Rather than seeking to maximize performance, we adopt a resource-oriented approach. By exploiting the Monte Carlo Dropout technique to estimate the predictive uncertainty of a link prediction model, we generate robust candidate triplets for potential integration into the resource.

The third part explores the opposite axis, "Resources for NLP", seeking to determine whether the integration of structural knowledge from a lexical-semantic network can improve lexical disambiguation. Chapter 5 presents a state-of-the-art overview of the lexical disambiguation task and an analysis of its traditional evaluation framework. In Chapter 6, we evaluate different lexical disambiguation systems, including the integration of graph embeddings, learned from `RL-fr`, into a supervised neural model.

In this thesis, we focused on the mutual enrichment between lexical resources and NLP systems. Although contemporary NLP techniques have evolved towards the use of large language models trained on large amounts of text, we emphasize the usefulness of lexical resources, due to their controlled and precise description of linguistic phenomena.

Publications

Conférences internationales

— **Beyond Model Performance : Can Link Prediction Enrich French Lexical Graphs ?**
Hee-Soo Choi, Priyansh Trivedi, Mathieu Constant, Karën Fort, Bruno Guillaume.
Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), mai 2024, Turin, Italie.
`https://aclanthology.org/2024.lrec-main.208/`

Conférences nationales

— **Désambiguïsation lexicale d'exemples lexicographiques du français : intégration de plongements de graphe et évaluation**
Hee-Soo Choi, Mathieu Constant, Karën Fort, Bruno Guillaume.
Actes de la 33e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), juillet 2026, Nantes, France.
`https://talnarchives.atala.org/TALN/TALN-2026/57.pdf`

— **Au-delà de la performance des modèles : la prédiction de liens peut-elle enrichir des graphes lexico-sémantiques du français ?**
Hee-Soo Choi, Priyansh Trivedi, Mathieu Constant, Karën Fort, Bruno Guillaume.
Actes de la 31e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), juillet 2024, Toulouse, France.
`https://aclanthology.org/2024.jeptalnrecital-taln.3/`

— **Des ressources lexicales du français et de leur utilisation en TAL : étude des actes de TALN.**
Hee-Soo Choi, Karën Fort, Bruno Guillaume, Mathieu Constant.
Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), juin 2023, Paris, France.
`https://aclanthology.org/2023.jeptalnrecital-short.3/`

— **État de l'art : Liage de ressources lexicales du français.**
Hee-Soo Choi.
Actes de la 24e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL), juin 2022, Avignon, France.
`https://aclanthology.org/2022.jeptalnrecital-recital.5/`

Introduction

L'étude du vocabulaire d'une langue est fondamentale à sa compréhension et à sa modélisation. En linguistique, l'analyse de la langue et de ce que l'on appelle communément ses « mots », passe par un travail de description, pouvant aboutir à la construction de ressources lexicales. Ces ressources constituent alors des traces écrites et formalisées du travail des linguistes et permettent d'encoder des connaissances relatives aux mots, à leurs propriétés morphologiques, syntaxiques, sémantiques ou encore pragmatiques.

Les ressources lexicales ont joué un rôle fondamental dès les débuts du traitement automatique des langues (TAL), par leur description explicite et structurée du vocabulaire, nécessaire à son traitement automatique. Historiquement, elles ont permis la conception des premiers systèmes de TAL, fondés sur des approches symboliques et à base de règles. Un exemple notable est `WordNet` (Miller, 1995), réseau lexico-sémantique de l'anglais issu de la psycholinguistique, devenue l'une des ressources les plus utilisées en TAL (Mariani et al., 2019a).

Cependant, toutes les langues n'ont pas bénéficié de la même disponibilité de ressources lexicales, ce qui a entraîné des disparités dans leur traitement automatique. Pour le français plus spécifiquement, la communauté francophone du TAL s'est attachée à pallier ce manque par la création de diverses ressources lexicales, allant de l'adaptation de ressources issues de la linguistique à des ressources conçues spécifiquement pour le TAL, permettant ainsi au français d'atteindre le statut de langue bien dotée.

Dans les années 2010, le domaine a connu une évolution majeure avec l'essor des approches neuronales (Mikolov et al., 2013) et des grands modèles de langue, *Large Language Models* (LLM) (Devlin et al., 2019; Brown et al., 2020). Ces approches reposent sur l'exploitation de grandes quantités de données, majoritairement issues du *web*, à partir desquelles les systèmes apprennent à inférer des régularités et des motifs linguistiques. La modélisation explicite et théorique de la langue laisse alors place à une modélisation probabiliste, fondée sur les fréquences d'apparition et les cooccurrences observées dans les données. Cette évolution a entraîné une diminution progressive de l'usage des ressources lexicales, au profit de l'apprentissage sur de grandes quantités de textes non structurés.

Si les performances des modèles neuronaux surpassent celles des approches symboliques et basées sur des règles dans de nombreuses tâches lexicales et sémantiques, ce changement de paradigme invite à reconsidérer la place des ressources lexicales dans le TAL contemporain et soulève plusieurs enjeux, tant pour le domaine que pour les ressources elles-mêmes.

1 Enjeu pour le TAL : comment et pourquoi représenter les connaissances linguistiques ?

Les modalités de représentation des connaissances linguistiques ont fortement évolué en TAL avec le passage de représentations symboliques, lisibles par l'humain et adaptables à la machine à des représentations numériques uniquement lisibles par la machine. Plusieurs questions se posent alors : dans quelle mesure ces représentations vectorielles parviennent-elles à encoder des connaissances linguistiques ? Permettent-elles d'encoder tous les phénomènes linguistiques ? De toutes les langues ?

Des représentations vectorielles peu transparentes. Les grands modèles de langue sont des réseaux de neurones profonds basés sur l'architecture transformeur (Vaswani et al., 2017) et entraînés sur de grandes quantités de textes grâce à des techniques d'apprentissage auto-supervisé (Rogers and Luccioni, 2024). Ils présentent aujourd'hui les meilleures performances sur des tâches lexicales et sémantiques telles que la traduction automatique (Barrault et al., 2019) ou la désambiguïsation lexicale (Bevilacqua et al., 2021). Ces systèmes résolvent des tâches en représentant des connaissances linguistiques de manière implicite, à travers des représentations vectorielles, appelées plongements (*embeddings*), apprises sur de grandes quantités de données. Plus précisément, les plongements sont appris durant l'entraînement de manière à capturer des régularités statistiques présentes dans les données.

Le caractère implicite de cette représentation pose notamment la question de la transparence et de l'interprétabilité de ces systèmes basés sur une architecture opaque. Des travaux ont démontré que les systèmes neuronaux permettent de coder un large éventail d'informations linguistiques à travers les couches de neurones (Blevins et al., 2018; Jawahar et al., 2019; Belinkov and Glass, 2019; Hewitt and Manning, 2019). Jawahar et al. (2019) montrent que les couches du modèle `BERT` (Devlin et al., 2019) codent une riche hiérarchie d'informations linguistiques, allant de caractéristiques de surface dans les premières couches, à des propriétés syntaxiques dans les couches intermédiaires, puis à des informations sémantiques dans les couches supérieures. Hewitt and Manning (2019) démontrent également que les modèles `BERT` et `ELMo` (Peters et al., 2018) encodent des arbres syntaxiques dans leurs plongements.

Bien que cette lignée de travaux s'inscrive dans une démarche visant à rendre ces nouvelles représentations interprétables, leur mise en œuvre s'avère plus complexe dans le cas des grands modèles de langue auto-régressifs, dont l'architecture composée de plusieurs milliards de paramètres est d'autant plus opaque (Singh et al., 2024). Avec l'omniprésence de ces systèmes dans le TAL contemporain, la question de la représentation des informations linguistiques reste largement ouverte.

Des phénomènes linguistiques faiblement représentés. La langue se caractérise par une grande richesse de phénomènes linguistiques, reflet de sa complexité et de sa diversité. Si les grands modèles de langue sont capables d'encoder certaines informations linguistiques, la question de savoir s'ils parviennent à capturer toute la finesse et la granularité de la langue à partir de grandes quantités de données demeure.

Fortement fondés sur des régularités statistiques et des fréquences d'apparition, les grands modèles de langue tendent à privilégier, dans leurs représentations internes, les motifs les plus

fréquemment observés dans les données d'entraînement (Bender et al., 2021; Kang and Choi, 2023; McCoy et al., 2023). Guo et al. (2025) montrent ainsi que, bien que les grands modèles de langue soient capables de produire des textes plausibles et cohérents, un écart significatif subsiste lorsqu'il s'agit de reproduire la richesse linguistique humaine dans des tâches créatives.

De même, Juzek and Ward (2025) observent une sur-utilisation de certains mots par `ChatGPT` [1], agent conversationnel basé sur le grand modèle de langue `GPT` (Brown et al., 2020). Toutefois, ils ne parviennent pas à déterminer si ce phénomène découle de l'architecture du modèle ou des données d'entraînement, en raison de leur inaccessibilité.

Des inégalités entre les langues. Les performances des outils de TAL varient fortement selon les langues considérées, avec un écart important entre langues très dotées et langues peu dotées (Bender, 2019; Joshi et al., 2020). Les systèmes actuels, dépendants de grandes quantités de données, rencontrent des difficultés majeures pour les langues qui n'en disposent pas. En effet, l'apprentissage de représentations vectorielles nécessite des corpus étendus, ce qui rend le traitement des langues peu dotées difficile, voire impossible. Cette pénurie de données accentue les disparités entre langues et limite la portée universelle des avancées en TAL (Joshi et al., 2020).

Malgré les capacités multilingues revendiquées par les grands modèles de langue modernes, les performances restent inégales. Les langues abondamment représentées dans les données du *web*, comme l'anglais, le chinois ou l'espagnol, bénéficient de modèles performants, tandis que les langues peu présentes, parmi lesquelles de nombreuses langues africaines ou asiatiques, sont sous-représentées, réduisant considérablement la qualité des représentations et des applications linguistiques (Hedderich et al., 2021).

2 Enjeu pour les ressources lexicales : comment valoriser leur caractère explicite ?

Bien que leur utilisation ait diminué dans le TAL contemporain, les ressources lexicales restent des sources de données de qualité, fournissant des informations détaillées sur le vocabulaire d'une langue. Dans ce contexte, les systèmes automatiques peuvent-ils exploiter pleinement ces informations ? Et à l'inverse, le TAL peut-il contribuer à enrichir et améliorer les ressources lexicales elles-mêmes ?

Intégrer des connaissances explicites dans les grands modèles de langue. L'intégration de connaissances symboliques dans les grands modèles de langue permet de pallier certaines limites des approches statistiques ou distributionnelles (Faruqui et al., 2015).

Ce processus contribue à répondre à la fois aux enjeux d'interprétabilité et à ceux liés à la sous-représentation de certains phénomènes linguistiques. Les connaissances symboliques peuvent se matérialiser, entre autres, sous la forme de données lexicales décrivant le vocabulaire d'une langue, ou de relations lexicales et sémantiques entre unités lexicales (Navigli and Ponzetto, 2012; Moskvoretskii et al., 2024). Les ressources lexicales peuvent également permettre le traitement des langues peu dotées, *via* l'utilisation de dictionnaires ou de grammaires descriptives (Zhang et al., 2024).

Il en résulte des systèmes hybrides, combinant des informations inférées à partir de vastes corpus textuels avec des données explicites issues des ressources lexicales.

1. `https://chatgpt.com/`

Des données utiles pour l'évaluation. Les ressources lexicales constituent des données de qualité, servant de référence pour l'évaluation des systèmes (Opitz et al., 2025). Elles offrent la possibilité de tester de manière systématique des phénomènes linguistiques tels que la polysémie, les relations sémantiques ou les variations morphologiques, souvent dilués dans des corpus de grande taille. Elles permettent ainsi de dépasser une évaluation globale fondée sur des métriques agrégées, en facilitant une analyse plus fine des capacités réelles des modèles et en mettant en évidence leurs lacunes face à certains phénomènes linguistiques rares ou complexes.

Par ailleurs, ces ressources peuvent être mobilisées dans une perspective de diagnostic, afin d'analyser les biais induits par les données d'entraînement ou les limites des représentations apprises. Elles peuvent contribuer à une meilleure compréhension du comportement des systèmes.

Enrichir et pérenniser les ressources lexicales. Le maintien et l'enrichissement des ressources lexicales constituent un enjeu fondamental dans un contexte où les langues évoluent continuellement, tant sur le plan lexical que sémantique. L'émergence de nouveaux usages, la spécialisation terminologique ou encore l'apparition de néologismes impliquent une actualisation régulière des descriptions linguistiques, afin de garantir leur pertinence pour les applications de TAL contemporaines.

Au-delà du TAL, les ressources lexicales constituent également des sources d'informations utiles dans des domaines comme l'analyse linguistique, la lexicographie, la traduction ou l'enseignement des langues. Elles se heurtent toutefois à la difficulté d'atteindre une couverture exhaustive en raison du caractère long et coûteux de leur construction (Gala, 2013).

Si les ressources lexicales peuvent bénéficier aux systèmes de TAL, les techniques automatiques peuvent également contribuer à compenser les limites liées à la construction et à l'enrichissement manuel des ressources lexicales (Sagot, 2010; Navigli and Ponzetto, 2012).

3 Cadre de la thèse

Questions de recherche. Les ressources lexicales et le TAL sont étroitement liés par une relation bilatérale. Cette thèse s'intéresse à cette interaction et vise à mettre en évidence l'apport réciproque que chacun peut offrir à l'autre.

Nous cherchons ainsi à traiter les enjeux pour le TAL et les ressources, soulevés précédemment, autour de cinq questions de recherche :

— QR 1 : comment a évolué la représentation de l'information linguistique dans les ressources lexicales, de la linguistique au TAL ?

— QR 2 : comment a évolué la présence des ressources lexicales du français sur une période de 25 ans d'articles scientifiques ? Quelles sont les ressources les plus utilisées et pour quelles applications ?

— QR 3 : un modèle de prédiction de liens peut-il prédire de potentiels nouveaux triplets pour un réseau lexico-sémantique du français ?

— QR 4 : l'intégration de plongements de graphe, appris à partir d'un réseau lexico-sémantique, peut-elle améliorer les performances d'un modèle de désambiguïsation lexicale du français ?

— QR 5 : le cadre d'évaluation de la tâche de désambiguïsation lexicale reflète-t-elle les performances réelles des modèles ?

Plan de la thèse. Cette thèse s'articule autour de trois parties et six chapitres.

La première partie est consacrée à l'évolution du concept de ressource lexicale, depuis sa définition en linguistique jusqu'à ses usages en TAL et répond aux deux premières questions de recherche (QR 1 et QR 2).

Le chapitre 1 est consacré à un état de l'art des ressources lexicales de la linguistique au TAL. Nous discutons de l'évolution de la place des ressources lexicales dans ce domaine, qui a été marqué par l'émergence des grands modèles de langue. En ce sens, une attention particulière est portée à la question de la représentation de les informations linguistiques avec le passage de représentations symboliques à des représentations numériques.

Le chapitre 2 présente une analyse bibliométrique des ressources lexicales du français dans les articles de la conférence nationale de TAL. Cette analyse nous permet d'observer statistiquement l'évolution de leur utilisation sur une période de 25 ans, et d'aborder des questions liées à la réutilisabilité.

Dans la deuxième partie, nous explorons le volet « TAL pour les ressources » avec l'enrichissement des réseaux lexico-sémantiques à travers des techniques de prédiction de liens. Nous traitons de la troisième question de recherche dans cette partie (QR 3).

La partie est introduite par le chapitre 3, dans lequel nous présentons un état de l'art sur la complétion de graphes et plus spécifiquement sur la tâche de prédiction de liens.

Le chapitre 4 décrit nos expériences sur l'enrichissement de réseaux lexico-sémantiques du français à partir d'un modèle de prédiction de liens. Contrairement à la majorité des travaux sur cette tâche, nous adoptons un point de vue davantage orienté vers la ressource et non le modèle, en interrogeant la capacité du modèle à fournir des informations exploitables à la complétion d'un réseau lexico-sémantique parcimonieux. La question centrale de cette partie est alors de savoir dans quelle mesure les sorties produites par les modèles peuvent constituer des informations linguistiquement pertinentes.

Enfin, la troisième partie aborde le second versant de la relation entre les ressources lexicales et le TAL, celui des « ressources pour le TAL ». Dans cette partie, nous traitons des deux dernières questions de recherche (QR 4 et QR 5). Plus précisément, nous cherchons à intégrer des informations linguistiques sous la forme de plongements de graphe appris à partir d'un réseau lexico-sémantique, dans un système de désambiguïsation lexicale du français. L'objectif est d'observer si la représentation des informations structurelles du graphe dans ces plongements permettent une amélioration du traitement de la polysémie, en particulier sur des données annotées avec une ressource de granularité fine.

Le chapitre 5 est ainsi consacré à un état de l'art de la tâche de désambiguïsation lexicale, une tâche historique et fondamentale en TAL, qui a été largement étudiée au cours des deux dernières décennies. Nous décrivons la tâche avec une attention particulière à l'évaluation.

Le chapitre 6 décrit ensuite les expériences d'intégration de plongements de graphe dans un système de désambiguïsation lexicale d'exemples lexicographiques du français. Nous présentons les résultats sur différents types de systèmes à travers une évaluation reflétant davantage les productions réelles des systèmes. Dans ce chapitre, nous cherchons à explorer si les plongements de graphe peuvent encoder la polysémie du français et si leur intégration dans un système de désambiguïsation lexicale peut améliorer ses performances.

Contexte et positionnement. Cette thèse a été menée dans un contexte interdisciplinaire, entre un laboratoire de linguistique, l'ATILF (Analyse et Traitement Informatique de la Langue Française) et un laboratoire d'informatique, le LORIA (Laboratoire Lorrain de Recherche en Informatique et ses Applications).

Ce contexte a influencé certains choix dans notre recherche, notamment sur la langue et les ressources étudiées. En effet, nos expériences exploitent uniquement des ressources lexicales du français, en particulier le `RL-fr` (Lux-Pogodalla and Polguère, 2011), réseau lexico-sémantique créé sous l'impulsion d'Alain Polguère, Professeur des universités à l'Université de Lorraine réalisant sa recherche à l'ATILF. Une partie de notre recherche s'intéressant à l'enrichissement d'une ressource, nous avons choisi le `RL-fr`, ressource créée entièrement manuellement, afin de fournir une contribution concrète à son amélioration.

Un deuxième choix a été de nous concentrer sur des réseaux lexico-sémantiques. Outre le `RL-fr`, nous avons également utilisé `RezoJDM` (Lafourcade and Joubert, 2008), un réseau lexico-sémantique du français de grande couverture créé par des mécanismes automatiques et à travers des jeux ayant un but. Le contraste avec le `RL-fr` nous permet de faire une comparaison de nos expériences sur deux réseaux lexico-sémantique très différents.

Enfin, dans cette thèse, nous consacrons une partie importante à la question de l'évaluation dans les expériences que nous menons, tant sur la tâche de prédiction de liens que la tâche de désambiguïsation lexicale.

La prédiction de liens cherche à répondre à la problématique d'incomplétude des réseaux lexico-sémantiques et plus généralement des graphes de connaissances. Bien que cet argument de l'incomplétude soit largement avancé dans les travaux de recherche, la configuration de la tâche se focalise davantage sur l'évaluation des performances des systèmes, qu'à l'enrichissement même de la ressource. Dans nos travaux, nous cherchons à aller au-delà des performances et à exploiter les sorties produites par le modèle, afin d'en extraire de nouvelles informations.

Concernant la tâche de désambiguïsation lexicale, nous proposons un état de l'art remettant en question le cadre d'évaluation, qui s'est ancré dans un schéma fixe depuis plusieurs années. Nous identifions ainsi différents biais provenant des données et des métriques, démontrant qu'une valeur numérique seule ne peut attester de la performance réelle d'un modèle. Dans nos expériences, nous fournissons ainsi les résultats des différents systèmes avec différentes métriques et nous conduisons une analyse qualitative, qui nous permet d'observer des phénomènes invisibles à travers une métrique seule.

Première partie

Les ressources lexicales en Traitement Automatique des Langues

Introduction

L'analyse et la modélisation de la langue reposent sur la nécessité de disposer d'une description de la langue, laquelle est notamment fournie par les ressources lexicales. En linguistique, celles-ci s'inscrivent généralement dans des cadres théoriques visant à rendre compte du fonctionnement de la langue de manière précise et cohérente. Les ressources lexicales peuvent ainsi être envisagées comme des traces du travail d'analyse mené par les linguistes et les lexicographes.

Avec le développement des méthodes automatiques et du traitement automatique des langues (TAL), ces ressources sont devenues des sources de connaissances linguistiques essentielles pour la conception de systèmes computationnels. Leur usage y est principalement applicatif plutôt que descriptif : les ressources lexicales constituent désormais des données d'entrée exploitées par les systèmes automatiques. Par ailleurs, les travaux en TAL se sont progressivement orientés vers la création de ressources de manière (semi-)automatique, afin de dépasser les limites d'une élaboration entièrement manuelle et de produire des ressources spécifiquement adaptées aux besoins du TAL.

Dans cette première partie, nous nous intéressons à l'évolution des ressources lexicales de la linguistique au TAL en présentant dans le chapitre 1, un état de l'art des ressources lexicales provenant de la linguistique et réutilisées ensuite en TAL. Dans ce chapitre, nous cherchons à répondre à la première question de recherche : comment a évolué la représentation de l'information linguistique dans les ressources lexicales, de la linguistique au TAL ?

Nous consacrons ainsi une partie importante à répondre à cette question de la représentation des connaissances linguistiques, celle-ci évoluant d'un format symbolique lisible par l'humain à un format numérique lisible uniquement par la machine. Nous ouvrons la discussion à la manière dont les grands modèles de langue, omniprésents dans le TAL actuel, encodent les informations lexicales.

Dans le chapitre 2, nous présentons une analyse bibliométrique de la dynamique des ressources lexicales du français dans la conférence nationale de TAL, TALN. Dans la même veine que les travaux de Mariani et al. (2019b), nous nous proposons de mesurer la présence des ressources lexicales sur une période de 25 ans, afin de répondre à notre deuxième question de recherche : comment a évolué la présence des ressources lexicales du français dans dans les articles scientifiques ? Nous abordons également des questions liées à la réutilisabilité des ressources, à la manière dont elles sont utilisées et pour quelles applications.

Chapitre 1

De la linguistique au traitement automatique des langues

Sommaire

1.1 Introduction 17
1.2 Qu'est-ce qu'une ressource lexicale ? 18
1.2.1 L'ambiguïté du mot 19
1.2.2 Type de ressource et contenu linguistique 19
1.2.3 Caractéristiques d'une ressource lexicale 21
1.3 Les ressources lexicales : empreintes des théories linguistiques . 22
1.3.1 Le Lexique-Grammaire et le distributionnalisme 22
1.3.2 Dicovalence : notion de valence et approche pronominale 23
1.3.3 Le Réseau lexical du français et la théorie Sens-Texte 24
1.3.4 La famille des *Net* : WordNet, FrameNet, VerbNet 25
1.3.5 L'influence de la linguistique générative 26
1.4 Les ressources lexicales : « matériau » pour le TAL 27
1.4.1 Briques fondamentales des approches symboliques et statistiques . . 27
1.4.2 Construction automatique : compromis entre couverture et qualité . 28
1.4.3 Pérennité et accessibilité des ressources lexicales 31
1.5 Les plongements lexicaux : une représentation neuronale des connaissances linguistiques 32
1.5.1 Plongements lexicaux statiques 32
1.5.2 Plongements lexicaux contextuels 33
1.6 La place des ressources lexicales aujourd'hui en TAL 33
1.6.1 Intégration de ressources lexicales aux grands modèles de langue . . 33
1.6.2 Enjeux pour les langues peu dotées 35
1.6.3 Les grands modèles de langue ont-ils des « compétences lexicales » ? 35
1.7 Conclusion 36

1.1 Introduction

Les ressources lexicales occupent une place centrale dans l'étude et la modélisation de la langue, qu'il s'agisse de linguistique ou de traitement automatique des langues (TAL). Dans ce

chapitre, nous nous proposons de décrire l'évolution des ressources lexicales à travers ces deux domaines, en termes de représentation des informations linguistiques et d'utilisation.

Pour ce faire, nous commençons par décrire ce qu'est une ressource lexicale dans la section 1.2 en détaillant la notion de mot et les caractéristiques propres à une ressource lexicale.

Avant d'être utilisées en TAL, les ressources lexicales ont permis d'illustrer les théories linguistiques qui modélisent la langue. Ces ressources constituent ainsi le « résultat » du travail des lexicographes et des linguistes. Nous en présentons un panorama dans la section 1.3, en mettant un accent sur la diversité des approches théoriques. Dans la section 1.4, nous nous concentrons sur les ressources lexicales depuis le TAL, qui ont constitué des briques fondamentales à la création des premiers systèmes automatiques. Le développement des techniques de TAL ont ensuite permis l'automatisation de la création de ressources davantage à l'image des besoins du domaine.

Dans la section 1.5, nous présentons l'évolution de la représentation des connaissances linguistiques avec le développement des grands modèles de langue neuronaux. Le vocabulaire d'une langue est alors représenté par des représentations numériques, les plongements lexicaux, remettant en question le rôle des ressources lexicales classiques. Nous ouvrons la discussion concernant cette question dans la section 1.6, où nous abordons les modèles hybrides, les enjeux pour les langues peu dotées et le devenir des ressources lexicales.

Nous précisons que nous nous concentrons majoritairement sur des ressources lexicales de l'anglais et du français dans ce chapitre.

1.2 Qu'est-ce qu'une ressource lexicale ?

Depuis les débuts de l'écriture, les ressources lexicales constituent des traces matérielles des langues, permettant ainsi leur description, leur préservation et leur transmission au fil du temps. La définition d'une ressource lexicale est large et englobe aujourd'hui une variété d'objets qui décrivent la langue sous différentes formes. Gala (2013) en donne la définition suivante :

> « [...] tout ouvrage - quel que soit son support - contenant des mots (unités lexicales, lexies, etc.) ou des concepts associés à des informations de nature très différente : des traductions vers une autre langue (ou vers d'autres langues) ou bien des explications à caractère linguistique (origine, caractéristiques grammaticales, emploi, etc.) ou conceptuel (liens thématiques, relations lexicales, etc.). »

Tout au long de cette thèse, nous retenons cette définition qui permet de désigner à la fois des ressources lexicales physiques provenant de la lexicographie et des ressources numériques créées et utilisées en TAL. Dans ce dernier domaine, on retrouve fréquemment le terme de « base de données lexicales », que nous considérons ici comme un synonyme de ressource lexicale.

Nous distinguerons également les ressources lexicales des ressources langagières, *language resources* en anglais. Les ressources langagières sont un ensemble plus large que les ressources lexicales dans le mesure où elles englobent les données décrivant la langue telles que des corpus écrits ou oraux, des bases de données terminologiques ou des recueil de discours, mais également les outils permettant d'améliorer ou d'évaluer les systèmes de traitement automatique des langues [2]. Les ressources lexicales représentent ainsi un sous-ensemble des ressources langagières qui se focalise sur la description du vocabulaire d'une langue.

2. `https://www.elra.info/about/what-is-a-language-resource/`

Dans cette section, nous définissons dans un premier temps la notion de « mot », unité élémentaire des ressources lexicales mais qui nécessite de fixer une terminologie précise. Nous détaillons ensuite les différents types de ressources lexicales en termes de format et d'informations qu'elles décrivent et les caractéristiques qui définissent une ressource lexicale.

1.2.1 L'ambiguïté du mot

Une ressource lexicale a pour unité minimale, une unité lexicale qui représente ce que l'on appelle communément un « mot ». En linguistique, la notion de mot a un caractère ambigu et ne présente pas de définition fixe, malgré son rôle central dans le domaine (Haspelmath, 2023). Elle est particulièrement discutable en raison de la variation entre les différentes langues, notamment sur les langues agglutinantes.

Pour le français, Polguère (2016) décrit les exemples suivants :

1. Sa réponse se tient en deux **mots** : « Sûrement pas ! »
2. « Suis », « es », « est », « sommes » ... sont des formes du même **mot**.

Dans le premier exemple, le terme est employé pour désigner deux signes linguistiques séparés à l'écrit par des espaces tandis que dans le deuxième exemple, il désigne une entité abstraite, plus générale qu'une forme linguistique.

Dans cette thèse, nous reprenons la terminologie de Polguère (2016) et d'Haspelmath (2023) qui distinguent les concepts de **mot-forme** et de **lexème**.

Le mot-forme correspond à une occurrence linguistique concrète, susceptible de fonctionner de manière autonome dans un énoncé, et caractérisée par des propriétés morphosyntaxiques spécifiques (temps, nombre, personne, etc.). Dans l'exemple 1, « Sûrement » et « pas » sont deux mots-formes.

Le lexème, quant à lui, est une unité lexicale abstraite qui regroupe l'ensemble des mots-formes partageant un même noyau de sens et appartenant à un même paradigme flexionnel. Ainsi, « être » dans le deuxième exemple, est un lexème.

1.2.2 Type de ressource et contenu linguistique

La description des unités lexicales varie fortement selon la nature des informations linguistiques que l'on cherche à représenter. En effet, le niveau de description linguistique, qu'il concerne la morphologie, la phonologie, la syntaxe ou la sémantique, influe directement sur la représentation formelle de la ressource et conduit à l'existence de différents types de ressources, notamment en termes de format.

Dans cette section, nous décrivons les objets liant une unité lexicale à des informations linguistiques, tels que les dictionnaires, les lexiques et les réseaux linguistiques. En ce sens, nous ne considérons pas les corpus bruts, les corpus annotés, les grammaires formelles et les bases de connaissances générales (connaissances du monde) et d'un domaine spécifique (médecine, finance, biologie etc.) qui relèvent davantage des ressources langagières que lexicales.

1.2.2.1 Dictionnaires et lexiques

Traditionnellement, le vocabulaire d'une langue est décrit sous la forme d'une liste de lexèmes triés par ordre alphabétique, constituant ainsi les **dictionnaires** définis comme suit :

> « (Lexicographie) Ouvrage de référence qui répertorie des mots dans un ordre convenu (alphabétique en général) fournissant pour chacun d'eux différentes informations concernant son sens ou son emploi. » [3]

Les principales informations présentes dans un dictionnaire sont généralement les définitions, les parties du discours, le genre ou des exemples d'utilisation. Outre les traditionnels dictionnaires Le Robert ou Le Larousse, un exemple célèbre de dictionnaire électronique est le Trésor de la Langue Française informatisé (`TLFi`) [4], version électronique du Trésor de la Langue Française, dictionnaire de la langue française des XIX^{e} siècle et XX^{e} siècle publié entre 1971 et 1994.

Un extrait de l'entrée « manger » dans le `TLFi` [5] est illustré ci-dessous :

> MANGER1, verbe trans.
>
> A. Avaler (un aliment solide ou pâteux) après (l')avoir mâché.
> 1. [Le compl. d'obj. est exprimé] *Les chevaux mangent du foin, de l'avoine (Ac.). Gaga et Clarisse faisaient avec Blanche un repas sérieux, mangeant des sandwiches sur une couverture étalée.* (ZOLA, Nana, 1880, p. 1388).
> *Les longues vaches flamandes (...) viennent en ronflant de plaisir manger parfois l'avoine au creux de sa main* (BERNANOS, M. Ouine, 1943, p. 1430) »

Un autre terme souvent employé pour décrire ces listes de lexèmes est le terme de **lexique**. Provenant du grec ancien *lexikón* signifiant « livre de mots », il désigne un ensemble de lexèmes d'une langue, d'un locuteur ou d'un domaine spécialisé. Par exemple, le `Le`***`fff`*** (Clément et al., 2004; Sagot, 2010) (Lexique des formes fléchies du français), est un lexique morphologique et syntaxique associant à chaque forme fléchie, correspondant à un mot-forme, des informations morphologiques (lexème, étiquette morphologique) et syntaxiques (cadre de sous-catégorisation, autres propriétés syntaxiques).

L'exemple ci-dessous décrit partiellement l'entrée « manger » dans le `Le`***`fff`*** :

```
mange v [pred='manger_____1<suj:sn|sinf|scompl,obj:(sn)>',cat=v,@imperative,@Y2s]
mange v [pred='manger_____1<suj:sn|sinf|scompl,obj:(sn)>',cat=v,@PS13s]
mangea  v [pred='manger_____1<suj:sn|sinf|scompl,obj:(sn)>',cat=v,@J3s]
mangeai  v [pred='manger_____1<suj:sn|sinf|scompl,obj:(sn)>',cat=v,@J1s]
```

Les dictionnaires et les lexiques décrivent ainsi des informations propres aux lexèmes, permettant une représentation sous forme de liste d'items. Ces informations peuvent relever de différents domaines linguistiques telles que la morphologie, la syntaxe, la phonologie et la sémantique.

Nous relevons toutefois que la distinction entre ces deux termes n'est pas entièrement claire. Si les dictionnaires contiennent traditionnellement des définitions, le terme est également utilisé pour des ressources qui n'en contiennent pas, comme `Dicovalence` (Van den Eynde and Mertens, 2006, 2010) dictionnaire qui présente des informations sur les verbes. De même, des lexiques peuvent également fournir des exemples d'usage comme le Lexique-Grammaire (Gross, 1975).

3. `https://fr.wiktionary.org/wiki/dictionnaire`
4. `https://www.cnrtl.fr/definition/`
5. Extrait du mot « MANGER » dans le TLFi - `https://www.cnrtl.fr/definition/manger`.

1.2.2.2 Les connaissances linguistiques sous forme de réseau

Si les listes des lexèmes d'une langue permettent d'associer des informations linguistiques propres à chacun, ce format s'avère peu adapté pour représenter les liens entre les lexèmes et les contextes dans lesquels ils apparaissent, contrairement à la représentation sous forme de réseau (Gala, 2013). Dans ce type de ressource, les lexèmes ne sont pas considérés de manière isolée mais en tant que nœuds interconnectés au sein d'un graphe.

Plus particulièrement, en sémantique, la représentation en réseau permet de décrire l'organisation du lexique mental et des régularités sémantiques de la langue. Un exemple notable est le `WordNet` (Miller, 1995), réseau lexico-sémantique de l'anglais, issu des sciences cognitives.

Les relations décrites dans un réseau lexico-sémantique peuvent être de deux types (Saussure et al., 1995) :

— Les **relations paradigmatiques**, telles que la synonymie, l'hyperonymie et l'hyponymie, la méronymie et l'holonymie, mais également des relations d'antonymie ou de dérivation morphologique. Elles lient les unités lexicales par des connexions sémantiques et/ou morphologiques.
En reprenant l'exemple de Polguère (2016), le lexème « barbe » est lié aux lexèmes « barbu », « barbiche » ou « bouc ».

— Les **relations syntagmatiques** lient les unités lexicales au niveau de la phrase selon des affinités combinatoires : « barbe » peut être lié aux adjectifs « grosse », « épaisse » ou aux verbes « se tailler », « se couper ».

Pour le français, on compte divers réseaux lexico-sémantiques dont le `WOLF` (Sagot and Fišer, 2008), version française et disponible librement de `WordNet`, le `RL-fr` (Lux-Pogodalla and Polguère, 2011) (cf. Section 1.3.3) et `RezoJDM` (Lafourcade and Joubert, 2008; Lafourcade and Le Brun, 2020) (cf. Section 1.4.2.2).

La morphologie peut également être représentée sous la forme d'un réseau, comme dans `Morphonette` (Hathout, 2008), créé à partir de `TLFnome` (Hathout et al., 2002).

1.2.3 Caractéristiques d'une ressource lexicale

La notion de ressource lexicale recouvre une réalité particulièrement large, en raison de la richesse des langues et de la diversité des cadres théoriques et méthodologiques permettant de les décrire. Cette diversité se manifeste par une hétérogénéité des ressources à plusieurs niveaux :

— le **cadre théorique linguistique** sur lequel repose la ressource ;

— le **niveau de description linguistique** considéré : morphologie, syntaxe, sémantique, phonologie ou traduction ;

— le **format de représentation** adopté : lexique, dictionnaire, réseau lexical, sémantique ou morphologique ;

— le **mode de construction** de la ressource : manuel, semi-automatique ou fondé sur des approches participatives (cf. Section 1.4.2.2) ;

— la **couverture linguistique**, en termes de nombre d'entrées ou de degré de granularité de l'information fournie.

Malgré cette variabilité, les ressources lexicales partagent la caractéristique de décrire la langue à un instant donné, sous une forme essentiellement fixe et statique, ce qui les rend intrinsèquement incomplètes. Si le développement des ressources lexicales informatisées a permis d'accroître significativement leur couverture et de faciliter leur mise à jour, l'élaboration d'une description exhaustive demeure néanmoins difficile, voire hors d'atteinte, face au caractère dynamique et évolutif des langues.

1.3 Les ressources lexicales : empreintes des théories linguistiques

La linguistique, depuis les fondements posés par Saussure (Saussure et al., 1995), n'a cessé d'évoluer à travers une diversité de cadres théoriques visant à modéliser le fonctionnement de la langue (Paveau and Sarfati, 2003). Ces théories, qu'elles soient structurales, distributionnelles ou sémantiques, se concrétisent dans la construction de ressources lexicales qui reflètent les choix méthodologiques et les postulats des modèles qui les sous-tendent.

Dans cette section, nous présentons des exemples de ressources lexicales qui illustrent différentes théories linguistiques et leurs approches de modélisation de la langue.

1.3.1 Le Lexique-Grammaire et le distributionnalisme

À la fin des années 60, inspiré du distributionnalisme de Harris, Maurice Gross applique la théorie au français en créant le `Lexique-Grammaire`, ressource conçue manuellement décrivant les propriétés syntaxiques de noms prédicatifs, verbes et adjectifs (Gross, 1975). L'approche de Gross est basée sur une description systématique des structures syntaxiques et sémantiques de la langue, s'opposant à l'idée que les principes théoriques sont capables d'unifier et d'induire des règles générales à partir d'informations éparses (Paveau and Sarfati, 2003). Plus précisément, il se repose sur la notion de transformation grammaticale qui permet de passer d'une structure syntaxique à une autre en conservant la sémantique. Par conséquent, l'unité minimale de sens considérée est la phrase minimale (sujet - verbe - complément), le sens d'un lexème étant fortement dépendant de ses contextes de distribution.

Par cette méthode d'analyse transformationnelle uniquement basée sur des critères formels, Gross aboutit à l'élaboration d'une classification, des verbes dans un premier temps. Il démontre le principe suivant : « il n'existe pas deux verbes ayant le même comportement distributionnel, mais les verbes peuvent être regroupés en classes ayant des analogies significatives de comportement, appelées classes d'équivalence » (Gross, 1975; Paveau and Sarfati, 2003).

À titre d'exemple, la transformation pronominale des verbes « obéir » et « penser » montrent qu'ils ne peuvent pas être regroupés dans la même classe :

Marie obéit à Pierre / Marie lui obéit
Marie pense à Pierre / *Marie lui pense

La ressource en elle-même est organisée en tables (ou classes), une table regroupant des éléments lexicaux d'une catégorie grammaticale partageant des propriétés. Un exemple d'une table est donnée en Figure 1.1. La table 33 représente la classe des verbes qui se construisent avec un argument introduit par la préposition « à » (Tolone, 2009). Chaque ligne correspond à une entrée lexicale, ici un verbe, et chaque colonne correspond à une propriété de la classe, qui peut être respectée ou non par chaque entrée (traduit par des signes +/- dans la matrice).

Une entrée lexicale ayant n sens est représentée par n entrées dans la table, comme le verbe « se rendre » dans l'exemple, les deux sens ne partageant pas les mêmes propriétés.

N0 =: Nhum	N0 =: N-hum	N0 =: Nnr	Ppv	Ppv =: se figé	Ppv =: en figé	Ppv =: les figé	Nég	<ENT>	N0 V	N0 être V-ant	N1 =: Nhum	N1 =: N-hum	N1 =: le fait Qu P	Ppv =: lui	Ppv =: y	N0hum V W sur ce point	[extrap]	<OPT>
+	-	-	<E>	-	-	-	-	renaître	+	+	-	+	-	-	+	-	-	Max renaît au bonheur de vivre
+	-	-	se	+	-	-	-	rendre	+	-	+	+	+	-	+	+	+	Max s'est rendu à mon opinion
+	-	-	se	+	-	-	-	rendre	+	-	+	-	-	-	-	-	-	Le caporal s'est rendu à l'ennemi
+	-	-	<E>	-	-	-	-	renoncer	-	-	+	+	-	-	+	-	-	Max renonce à son héritage

FIGURE 1.1 – Extrait de la table 33 des verbes (Figure 1 de Tolone (2009)).

Bien qu'initialement non conçue pour le TAL, la ressource a reçu un fort intérêt des talistes notamment pour son exigence de formalisation permettant une application potentielle au TAL. De nombreux travaux ont exploré la traduction des tables du Lexique-Grammaire en un format adapté (Gardent et al., 2005; Tolone, 2009; Laporte, 2010).

1.3.2 Dicovalence : notion de valence et approche pronominale

L'approche pronominale développée par Claire Blanche-Benveniste propose une analyse de la syntaxe du français centrée sur le rôle structurant des pronoms clitiques, considérés comme révélateurs de la valence des verbes. Contrairement aux descriptions traditionnelles fondées sur les groupes nominaux « pleins », cette approche montre que ce sont les pronoms (le, lui, en, y, etc.) qui indiquent de manière stable quels compléments un verbe peut ou doit prendre (Blanche-Benveniste, 2013). Prenons l'exemple du verbe « donner » :

Je **le lui** donne.

Le verbe « donner » se manifeste avec trois pronoms, ce qui met en évidence sa valence à trois actants : donneur, objet donné, bénéficiaire. L'approche pronominale met ainsi l'accent sur l'usage réel, notamment à l'oral, et fournit un outil empirique pour décrire les structures verbales du français.

La ressource `Dicovalence` [6] s'inscrit dans cette perspective et constitue un dictionnaire électronique de valence verbale du français (Van den Eynde and Mertens, 2006, 2010). Le dictionnaire décrit les verbes à partir de leurs schémas de construction attestés, en précisant les compléments possibles, leur réalisation syntaxique et surtout les pronoms clitiques correspondants (cf. Figure 1.2). Chaque entrée verbale liste ainsi les cadres de valence (par exemple transitif direct, indirect, à complément prépositionnel) et les contraintes associées.

6. `https://www.ortolang.fr/market/lexicons/dicovalence`

```
VAL$ donner: P0 P1 (P2) | AdjunctVerb
VTYPE$ adjunct_verb:3 augmenting
VERB$ DONNER/donner
NUM$ 31575
EG$ il m'a donné son devoir à corriger
TR_DU$ ?geven
TR_EN$ give to
FRAME$ subj:pron|n:[hum], obj:pron|n:[abs], ?objà:pron|n:[hum,nhum,?abs]
P0$ qui, je, nous, elle, il, ils, on, celui-ci, ceux-ci
P1$ que, la, le, les, en Q, ça, celui-ci, ceux-ci
P2$ 0, quoi, qui, me, lui, leur, ça, ceci, celui-ci, ceux-ci
AUX$ avoir
```

FIGURE 1.2 – Exemple du verbe « donner » dans Dicovalence.

1.3.3 Le Réseau lexical du français et la théorie Sens-Texte

Vers la fin des années 80, Igor Mel'čuk formalise la théorie Sens-Texte, qui considère la langue comme un système relationnel entre le contenu / le sens (la sémantique) d'un énoncé et sa forme / son texte (la phonétique) (Mel'čuk, 1996; Mel'čuk et al., 2012).

Contrairement au distributionnalisme et à l'approche pronominale qui adoptent un point de vue syntaxique, la théorie Sens-Texte adopte un point de vue principalement axé sur la sémantique, bien qu'elle décrit également des niveaux de représentation syntaxique et morphologique. Plus spécifiquement, le modèle sens-texte est composé de sept niveaux de représentation : sémantique (SemR), syntaxique profond (DSyntR), syntaxique de surface (SSyntR), morphologique profonde (DMorphR), morphologique de surface (SMorphR), phonétique profond (DPhonR) et phonétique de surface (SPhonR) (cf. Figure 1.3).

$$\text{SemR} \underset{\textit{Analyse}}{\longleftarrow} \text{DSyntR} \longleftrightarrow \text{SSyntR} \longleftrightarrow \text{DMorphR} \longleftrightarrow \text{SMorphR} \longleftrightarrow \text{DPhonR} \overset{\textit{Génération}}{\longrightarrow} \text{SPhonR}$$

FIGURE 1.3 – Schéma du modèle sens-texte et des structures composant les représentations sémantiques et syntaxiques profondes (Figure 1 de Cousin (2023)).

Si de nombreuses théories contemporaines supposent des niveaux syntaxiques et morphologiques, la théorie Sens-Texte les considère comme des niveaux intermédiaires entre le niveau sémantique (le sens) et le niveau phonétique (le texte). Ainsi, la correspondance entre les sens et les textes est entièrement modulaire : une correspondance entre les niveaux sémantique et syntaxique profond, une correspondance entre les niveaux syntaxique profond et syntaxique superficiel, une correspondance entre les niveaux syntaxique superficiel et morphologique profond, etc (Kahane, 2003).

Le `RL-fr` est un réseau lexico-sémantique du français, créé manuellement par des lexicographes et basé sur les principes de la théorie Sens-Texte (Lux-Pogodalla and Polguère, 2011). Les liens entre les unités lexicales sont décrits par des fonctions lexicales qui encodent des relations paradigmatiques et syntagmatiques. Quelques exemples de fonctions lexicales paradigmatiques

sont donnés ci-dessous :

— Synonymie (Syn) : `vélo` → `bicyclette`

— Antonymie (Anti) : `accord` → `désaccord`

— Hyperonymie (Gener) : `amour` → `sentiment`

Les relations syntagmatiques comprennent, entre autres, les collocations et les verbes supports :

— Intensifier (Magn) : `boire` → `comme un trou`

— Verbe support (Oper) : `danger` → `courir`

Le `RL-fr` illustre ainsi la correspondance entre le sens et le texte à travers les fonctions lexicales, représentant les dépendances sémantiques et syntaxiques entre les unités lexicales.

1.3.4 La famille des *Net* : WordNet, FrameNet, VerbNet

`WordNet` (Miller, 1995), `FrameNet` (Baker et al., 1998) et `VerbNet` (Kipper-Schuler, 2005) sont trois ressources lexicales fondamentales basées sur différents théories linguistiques. Les représentations sémantiques de ces ressources ont largement été utilisées en TAL, individuellement mais également de manière complémentaire (Baker and Fellbaum, 2009), notamment dans des tâches d'étiquetage des rôles sémantiques, d'extraction d'information et de traduction automatique (Baker, 2014; Danlos et al., 2016).

WordNet, une ressource issue des sciences cognitives. `WordNet` est un réseau lexical à grande échelle organisé autour de la sémantique lexicale. Les mots sont regroupés en ensembles de synonymes cognitifs appelés *synsets*, chacun représentant un concept distinct (Miller, 1995). Les *synsets* sont reliés entre eux par des relations sémantiques comme l'hyperonymie/hyponymie (cf. Figure 1.4).

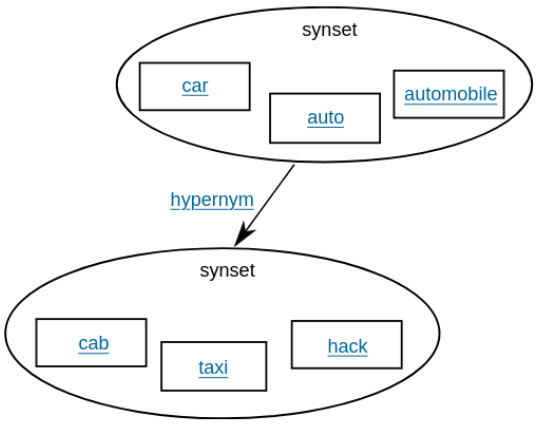


FIGURE 1.4 – Exemple de relation entre *synsets* (Figure de la documentation WordNet [7]).

La ressource est initialement conçue dans le but de représenter la sémantique structurelle et cognitive, c'est-à-dire la manière dont les humains organisent mentalement le sens des mots.

7. `https://wn.readthedocs.io/en/latest/guides/wordnet.html`

Son format numérique et son réseau sémantique riche le rendent particulièrement utile pour des tâches telles que la désambiguïsation des mots, le calcul de similarité sémantique, la recherche d'informations et la construction d'ontologies.

FrameNet et la sémantique des cadres. La ressource `FrameNet` (Baker et al., 1998) s'appuie sur la sémantique des cadres, une théorie qui postule que le sens des mots dépend de situations structurées ou des cadres dans lesquels ils apparaissent (Fillmore, 1982).

Elle repose sur la notion de cadre sémantique, une structure conceptuelle qui décrit une situation, un événement ou un état, ainsi que les participants typiques associés (appelés éléments du cadre).

Chaque unité lexicale, souvent un verbe, mais aussi des noms ou adjectifs, évoque un cadre particulier et contraint les rôles sémantiques de ses arguments. La phrase suivante présente un exemple du verbe « buy » :

Abby bought a car from Robin for $ 5,000. [8]

Le verbe « buy » évoque le cadre `Commerce_buy`, qui comprend des rôles comme *Buyer* (acheteur), *Seller* (vendeur), *Goods* (bien acheté) et *Money* (argent). « Abby » remplit le rôle d'acheteur, « car » celui de bien, « Robin » celui de vendeur et « $ 5,000 » celui d'argent.

VerbNet et les classes de Levin. `VerbNet` est une ressource lexicale computationnelle dédiée à la description systématique des verbes de l'anglais, organisée en classes sémantico-syntaxiques (Kipper-Schuler, 2005). Elle repose sur la théorie des classes verbales proposée par Levin, selon laquelle les verbes partageant des propriétés sémantiques manifestent des comportements syntaxiques similaires (Levin, 1993).

Sur la base de cette hypothèse, `VerbNet` regroupe les verbes en classes et sous-classes homogènes, chacune associée à un ensemble de cadres syntaxiques, de rôles thématiques (Agent, Theme, Patient, Instrument etc.) et de prédicats sémantiques formalisés. Par exemple, la classe `Break-45.1` [9] regroupe des verbes exprimant un changement d'état résultatif, tels que *break*, *shatter* ou *crack*.

De même, la classe `Give-13.1` [10] modélise les verbes de transfert en intégrant l'alternance dative (give a book to Mary / give Mary a book).

1.3.5 L'influence de la linguistique générative

Le passage de la linguistique au TAL a été fortement influencé par la linguistique générative, théorisée par Noam Chomsky, qui introduit une vision formelle et basée sur des règles du langage (Chomsky, 1965, 2002). Le langage est considéré comme un système structuré capable de générer une infinité de phrases à partir d'un ensemble fini de règles. Cette vision a permis aux débuts de la linguistique computationnelle de représenter la langue à l'aide de grammaires formelles.

Les grammaires formelles ne sont pas des ressources lexicales, car elles décrivent des structures syntaxiques abstraites plutôt que de stocker des informations sur des éléments lexicaux

8. Exemple issu du cadre `Commerce_buy` : `https://framenet.icsi.berkeley.edu/frames/Commerce_buy`
9. `https://verbs.colorado.edu/verb-index/vn/break-45.1.php`
10. `https://verbs.colorado.edu/verb-index/vn/give-13.1.php`

individuels. Cependant, pour mettre en œuvre les idées de la grammaire générative de Chomsky dans un système de TAL effectif, il nécessite une description fine des unités lexicales la langue, notamment de leurs catégories grammaticales, des cadres de sous-catégorisation et des rôles sémantiques. Les ressources lexicales au sens où on l'entend ici ont donc été essentielles dans des systèmes comme `XTAG` (Doran et al., 1994) dans lequel les structures de la grammaire (des arbres `TAG` dans ce cas) sont systématiquement lexicalisées.

1.4 Les ressources lexicales : « matériau » pour le TAL

Si les ressources lexicales constituent des traces écrites permettant d'exemplifier des théories linguistiques, elles sont devenues les sources principales d'informations linguistiques pour le TAL. Elles ont ainsi évolué d'un rôle d'exemples illustrant des théories linguistiques vers celui de données de base fournies en entrée aux systèmes automatiques.

Dans cette section, nous décrivons dans un premier temps la manière dont les ressources lexicales ont été exploitées, notamment dans les premières approches symboliques et statistiques.

Les techniques de TAL ont permis de répondre à la problématique de la couverture des ressources lexicales, dont la conception repose traditionnellement sur des processus manuels, coûteux et chronophage. Dans un second temps, nous nous intéressons plus particulièrement aux ressources construites de manière semi-automatique à partir de méthodes de TAL, lesquelles visent à produire des ressources spécifiquement adaptées aux besoins du TAL.

Enfin, nous abordons les enjeux liés à la pérennité et à l'accessibilité des ressources dans une dernière section.

1.4.1 Briques fondamentales des approches symboliques et statistiques

Dans les débuts du TAL, les approches symboliques reposaient largement sur l'exploitation de ressources lexicales afin de modéliser explicitement la sémantique et la structure des langues. Des ressources telles que `WordNet` ont joué un rôle central dans de nombreuses tâches fondamentales, notamment l'étude de la similarité sémantique (Resnik, 1995) ou la désambiguïsation lexicale (Lesk, 1986).

Par ailleurs, les ressources lexicales ont été mobilisées dans des systèmes pour différentes tâches syntaxiques et morphologiques comme l'étiquetage morpho-syntaxique (Denis and Sagot, 2012) ou l'analyse syntaxique (Riezler et al., 2002).

Avec l'essor des approches statistiques à la fin des années 1990, les ressources lexicales ont continué à jouer un rôle-clé, souvent au sein de techniques hybrides combinant règles linguistiques et modèles probabilistes pour l'analyse syntaxique (Collins, 1997; Riezler et al., 2002), l'analyse de dépendances (Agirre et al., 2011) ou la traduction automatique (Lü et al., 2007), entre autres.

Cette période a également été marquée par un effort important de la communauté pour convertir des ressources lexicales initialement conçues pour la linguistique descriptive en formats électroniques exploitables par les systèmes de TAL. En effet, le succès de `WordNet` s'explique en grande partie par sa disponibilité sous une forme électronique (*machine-readable*), contrairement à de nombreux dictionnaires traditionnels, dont la numérisation et la structuration ont nécessité un travail conséquent (Ide et al., 2003).

Dans le contexte francophone, des travaux comme ceux de Mertens (2010) illustrent ces efforts de conversion et d'adaptation de ressources lexicales existantes, telles que `Dicovalence`, pour des

usages computationnels. Dans la même veine, Hadouche and Lapalme (2010) convertissent le dictionnaire Les Verbes français, `LVF`, dans un format XML, permettant leur exploitation dans des systèmes d'annotation syntactico-sémantique de corpus ou de construction d'ontologies (Sabatier and Le Pesant, 2013).

En outre, le TAL a progressivement contribué à l'automatisation partielle de la construction de ressources lexicales, afin de pallier le caractère chronophage de leur élaboration manuelle (cf. Section 1.4.2). Cette dynamique marque une transition vers la création de ressources explicitement conçues pour les besoins du TAL, plutôt que pour la seule description linguistique.

1.4.2 Construction automatique : compromis entre couverture et qualité

La construction de ressources lexicales repose sur une longue tradition d'une description manuelle par des linguistes et des lexicographes. Cependant, ce processus se trouve être généralement long et coûteux et ne favorise pas une large couverture, bien que la qualité du travail est préservée. Avec le développement de l'informatique et du TAL, de nouvelles méthodes de construction apparaissent, à la fois pour automatiser la phase de construction mais également pour répondre à nouveaux besoins.

Nous distinguons trois types de méthodes de construction (semi-)automatique (Choi, 2022) :
— la construction par fusion de ressources existantes,
— la construction par myriadisation (ou collaborative),
— la construction par traduction.

La figure 1.5 présente une partie des ressources lexicales du français disponibles pour le TAL selon leur méthode de construction.

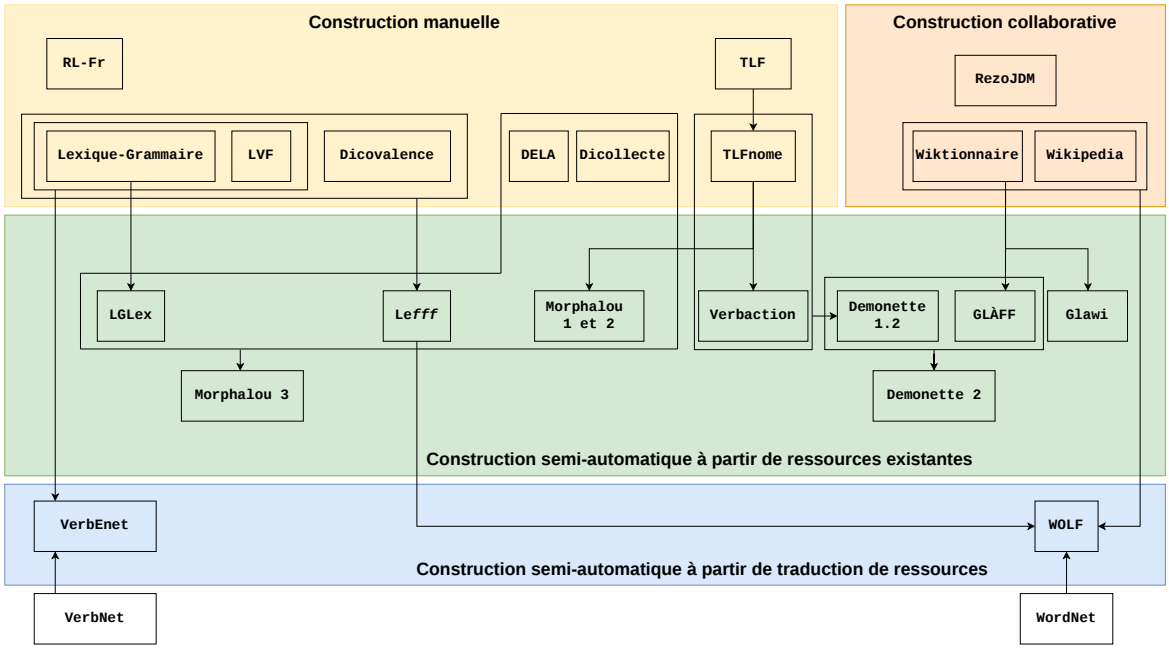


FIGURE 1.5 – Ressources lexicales du français selon la méthode de construction.

1.4.2.1 Construction par fusion

La richesse qui découle de l'hétérogénéité des ressources peut être exploitée dans la mesure où celles-ci présentent des données différentes et généralement complémentaires. Les rassembler permet ainsi leur enrichissement mutuel. Nous présentons trois exemples de ressources du français créés par fusion de ressources existantes : le Le*fff*, les différentes versions de Morphalou et Demonette.

Le Le*fff* (Sagot, 2010), le Lexique des Formes Fléchies du Français, est un large lexique flexionnel issu des tables du Lexique-Grammaire (Gross, 1975), du LVF (Dubois and Dubois-Charlier, 1997) et de Dicovalence (Van den Eynde and Mertens, 2006; Mertens, 2010) mais également de l'acquisition automatique avec validation manuelle de données sur corpus et d'informations syntaxiques.

Morphalou 3 (ATILF, 2019) est également un lexique des formes fléchies du français qui résulte de la fusion de cinq ressources : Morphalou 2, le Le*fff*, Dela[11], Dicollecte et LGLex. Initialement, Morphalou est issu du TLFnome (Hathout et al., 2002), la nomenclature du Trésor de la Langue Française et Morphalou 2 est une version contenant 30 000 lexèmes supplémentaires provenant du TLFi.

Si les deux ressources traitent de la morphologie, le Le*fff* couvre en plus des informations syntaxiques, tandis que Morphalou propose des transcriptions phonétiques d'une partie des lemmes et des formes fléchies.

Quant à Demonette (Hathout and Namer, 2014), base lexicale morpho-sémantique du français organisée en réseau dérivationnel, la version 1.2 provient du TLFnome et de Verbaction (Tanguy and Hathout, 2002), lui-même créé à partir du TLFnome. Demonette 2 a ensuite été construit à partir de Demonette 1.2, du GLÀFF (Sajous et al., 2013), Un Gros Lexique À tout Faire du Français et de ressources dérivationnelles développées et validées par des morphologues (Namer et al., 2019).

Bien qu'elles ne soient pas issues de la fusion de différentes ressources, certaines ressources telles que GLÀFF et Glawi (Hathout and Sajous, 2016) constituent également des adaptations d'une ressource existante, le Wiktionnaire, Wiktionary du français[12], en un format structuré plus accessible pour des travaux en TAL.

1.4.2.2 Construction par myriadisation

Avec le développement de l'informatique, l'approche collaborative ou participative (*crowdsourcing*), appelée aussi peuplonomie ou myriadisation (Sagot et al., 2011), permet de faire participer des personnes non expertes, c'est-à-dire sans formation linguistique ou lexicographique à proprement parler, mais intéressées par la langue.

Le principe est de solliciter un grand nombre de personnes pour effectuer de petites tâches difficilement automatisables (Lafourcade and Joubert, 2013). La contribution de la foule peut être rémunérée ou non, comme par exemple dans Wikipedia[13], ressource phare issue d'une construction collaborative bénévole. Quant à l'approche collaborative rémunérée à la tâche, elle

11. https://infolingu.univ-eiffel.fr//DonneesLinguistiques/Dictionnaires/telechargement.html
12. https://www.wiktionary.org/
13. https://www.wikipedia.org/

fait intervenir des problématiques éthiques qui ont été soulevées au sujet du système Amazon Mechanical Turk (Fort et al., 2011).

Si la participation à l'élaboration de `Wikipedia` et ses ressources connexes est motivée par l'intérêt de la foule sur un sujet particulier, il est également possible de la faire participer à travers des jeux ayant des buts. C'est le cas notamment de `RezoJDM` (Lafourcade and Joubert, 2008; Lafourcade and Le Brun, 2020), ressource lexico-sémantique créée à partir de la plate-forme `JeuxDeMots` [14] qui présente divers jeux en ligne. L'avantage de cette méthode est d'obtenir une grande quantité de données de manière rapide, peu coûteuse et dynamique mais questionne la qualité des données recueillies.

1.4.2.3 Construction par traduction

Les ressources précédemment présentées ont été créées à partir de ressources construites manuellement, tant par des experts que par la communauté de manière collaborative. D'autres ressources ont été construites de la même manière en ajoutant des traductions de ressources dans une autre langue. Nous pouvons citer les exemples de `Verb∋Net` (Danlos et al., 2014), `WOLF` (Sagot and Fišer, 2008) et du `French FrameNet` (Candito et al., 2014).

`Verb∋Net` a été créé à partir de `VerbNet` (Kipper-Schuler, 2005), un lexique inspiré des classes de Levin (Levin, 1993) qui décrit les caractéristiques syntaxiques et sémantiques des verbes de l'anglais (cf. Section 1.3.4). Pour le français, le `Lexique-Grammaire` et le `LVF` ont été utilisés pour faire correspondre leurs classes à celles de `VerbNet`. Danlos et al. (2014) ont tenu à conserver au maximum la structure des 270 classes de `VerbNet`. Ainsi, pour préserver la qualité des données, la ressource a été majoritairement construite manuellement avec une phase de traduction des verbes anglais de `VerbNet` *via* des dictionnaires bilingues, `Wiktionnaire` et `SCI-FRAN-EURADIC` [15].

Dans le même esprit, `WOLF`, WOrdNet Libre du Français, a été créé sur la base de `WordNet`, de ressources multilingues et de lexiques bilingues extraits de `Wikipedia` et `Wiktionary`. La ressource a été conçue afin de fournir un `WordNet` du français libre, contrairement au `French WordNet`, issu du projet `EuroWordNet` Vossen (1998) qui n'a été que très peu exploité en raison de sa licence restrictive. Dans `WOLF`, Sagot and Fišer (2008) adoptent deux approches : une approche de traduction et une approche d'alignement.

L'approche par traduction consiste à traduire les entrées monosémiques du `WordNet` en utilisant `Wikipedia` et ses ressources associées (`Wiktionary` anglais et français, `Wikispecies`) et le thesaurus multilingue `Eurovoc` [16].

Quant à l'approche d'alignement, elle permet de traiter les entrées polysémiques en les alignant sur un corpus parallèle en cinq langues (le français, l'anglais, le bulgare, le roumain et le tchèque) de `JRC-Acquis` (Steinberger et al., 2006). L'idée est d'extraire des informations sémantiques des traductions pour ensuite lever l'ambiguïté des entrées grâce aux `WordNets` de chaque langue. Quantitativement, cette méthode a permis au `WOLF` de couvrir davantage de *synsets* (32 351) que le `French WordNet` (22 121).

Le traitement des adverbes a également fait l'objet d'un enrichissement à partir de `DicoSyn` [17], un dictionnaire de synonymes, et du `Le`***`fff`*** (Sagot, 2010).

14. `http://jeuxdemots.org/`
15. `https://catalogue.elra.info/en-us/repository/browse/ELRA-M0033/`
16. `https://eur-lex.europa.eu/browse/eurovoc.html?locale=fr`
17. `https://crisco2.unicaen.fr/des/`

Quant au `French FrameNet`, il consiste en une version française du `FrameNet` (Baker et al., 1998), une ressource lexicale reposant sur la sémantique des cadres. Le `French FrameNet` vise à associer les entrées lexicales aux cadres de `FrameNet` et à fournir une couche d'annotations sémantiques sur le `French Treebank` (Abeillé et al., 2003; Abeillé and Barrier, 2004) et le corpus `Sequoia` (Candito and Seddah, 2012).

En outre, nous pouvons également citer `BabelNet` (Navigli and Ponzetto, 2012), réseau sémantique multilingue contenant du français, créé automatiquement à partir de `Wikipedia` et `WordNet`. Inspiré des `WordNets`, `BabelNet` regroupe des *Babel synsets* en plusieurs langues auxquels sont associées des définitions. Des techniques de traduction automatique sont également appliquées pour les entrées dans les langues peu dotées.

1.4.3 Pérennité et accessibilité des ressources lexicales

Le développement de ressources lexicales repose sur un processus long et coûteux, qu'il soit mené manuellement ou (semi-)automatique, dans la mesure où il nécessite des vérifications et validations régulières en cas d'erreurs ou de mises à jour. Dans ce contexte, assurer l'accessibilité et la réutilisabilité de ces ressource constitue un enjeu majeur, en particulier en TAL, où la disponibilité de données de qualité est essentielle.

Sagot (2013) souligne l'importance d'une distribution libre des ressources, dans le respect des droits des créateurs tels que définis par les licences associées, afin de permettre une meilleure exploitation et redistribution. Cet accès libre permet de surmonter les trois défis principaux que sont :

— la **visibilité**, qui constitue une condition préalable à une disponibilité à la fois immédiate et pérenne des ressources ;

— la **disponibilité immédiate**, nécessaire pour qu'une ressource puisse être citée, utilisée et éventuellement enrichie, et indispensable à la reproductibilité des résultats scientifiques ;

— la **disponibilité pérenne**, qui garantit l'exploitation des ressources sur le long terme, y compris après l'arrêt de leur maintenance par les développeurs initiaux, grâce à leur redistribution par des tiers.

La communauté du TAL s'est efforcée d'aller en ce sens avec la mise en place de plateformes permettant de distribuer les ressources, telles que la LRE Map [18], initiée par l'association européenne des ressources langagière (ELRA) et ORTOLANG [19], infrastructure nationale hébergeant des données et outils pour la valorisation des données sur le français et les langues de France.

Nous observons que plusieurs ressources présentées dans la section 1.4 ont été construites à partir de ressources pré-existantes, un processus rendu possible par l'accessibilité des données. Cette réutilisation a par ailleurs contribué de manière significative au développement des systèmes pour répondre à différentes tâches de TAL. Le tableau 2.1 présente une vue d'ensemble des ressources que nous avons décrites dans les sections 1.3 et 1.4 ainsi que leurs caractéristiques majeures.

Dans le chapitre 2, nous nous attacherons à analyser le degré de réutilisation de certaines ressources ainsi que les types d'applications dans lesquelles elles ont été mobilisées.

18. `https://lremap.elra.info/`
19. `https://www.ortolang.fr/fr/accueil/charte-ortolang/`

Nom	Langue	Domaine linguistique	Nombre d'entrées	Construction
`WordNet`	EN	Sémantique	117 659 *synsets*	Manuelle
`FrameNet`	EN	Syntaxe/Sémantique	10 462 unités lexicales	Manuelle
`VerbNet`	EN	Syntaxe/Sémantique	4 570 verbes	Manuelle
`BabelNet`	Multilingue	Sémantique	22 892 310 *synsets*	Automatique
`TLFi`	FR	Syntaxe/Sémantique	100 000 lexèmes	Manuelle
`Le`*`fff`*	FR	Syntaxe/Morphologie	536 375 formes fléchies	Semi-automatique
`Lexique-Grammaire`	FR	Syntaxe	∼75 000 lexèmes	Manuelle
`DELA`	FR	Morphologie	185 677 lexèmes	Manuelle
`Morphalou 3`	FR	Morphologie	93 681 lexèmes	Automatique
`Dicovalence`	FR	Syntaxe	∼3 700 verbes	Manuelle
`RezoJDM`	FR/EN	Sémantique/Syntaxe	15 616 619 termes	Automatique/Myriadisation
`RL-fr`	FR	Sémantique	29 975 unités lexicales	Manuelle
`WOLF`	FR	Sémantique	117 658 *synsets*	Automatique
`Verb∋Net`	FR	Syntaxe/Sémantique	∼3 000 verbes	Semi-automatique

TABLE 1.1 – Ressources mentionnées dans le chapitre 1 et librement disponibles.

1.5 Les plongements lexicaux : une représentation neuronale des connaissances linguistiques

L'arrivée des plongements lexicaux a transformé le TAL en introduisant des représentations continues du sens lexical, en rupture avec les approches symboliques et discrètes traditionnelles.

Avant leur développement, les lexèmes étaient généralement représentés par des identifiants ou des vecteurs creux de type *one-hot*, incapables de capturer les relations sémantiques ou syntaxiques entre termes. Les plongements lexicaux ont permis d'encoder ces relations directement dans l'espace vectoriel, où la similarité géométrique reflète la proximité de sens.

1.5.1 Plongements lexicaux statiques

Les plongements lexicaux statiques **représentent chaque lexème par un vecteur fixe**, indépendamment du contexte dans lequel il apparaît. Ils sont appris à partir de grands corpus de textes en exploitant les cooccurrences entre lexèmes.

`Word2Vec` (Mikolov et al., 2013) repose sur des réseaux de neurones peu profonds entraînés selon deux architectures principales : CBOW pour *Continuous Bag Of Words*, qui permet de prédire un lexème à partir de son contexte et Skip-gram, qui prédit le contexte à partir d'un lexème.

`GloVe` (*Global Vectors*) (Pennington et al., 2014) combine des approches statistiques et neuronales en factorisant une matrice globale de cooccurrences lexème-lexème, ce qui permet de capturer des régularités sémantiques à l'échelle du corpus entier.

Enfin, Bojanowski et al. (2017) étendent `word2Vec` avec les plongements lexicaux statiques `fastText`, qui représentent les lexèmes comme une composition de sous-lexèmes (n-grammes de caractères), ce qui améliore la gestion des lexèmes rares ou inconnus et des langues morphologiquement riches.

Avec ces modèles, un lexème donné correspond toujours au même vecteur, quelle que soit sa phrase d'apparition. Ainsi, on a une correspondance entre une représentation neuronale d'un lexème, comme nous disposons d'une correspondance entre un lexème et une définition dans

un dictionnaire. En ce sens, les plongements lexicaux statiques s'apparentent à une ressource lexicale bien que l'information linguistique décrite n'est pas sous forme textuelle, mais sous forme numérique.

1.5.2 Plongements lexicaux contextuels

Les plongements lexicaux contextuels produisent **une représentation différente d'un lexème selon son contexte**. Les modèles comme `BERT` (Devlin et al., 2019) s'appuient sur des architectures de type transformeur et sont pré-entraînés sur des tâches auto-supervisées, notamment le *Masked Language Modeling*, où le modèle apprend à prédire des lexèmes masqués en tenant compte à la fois du contexte gauche et droit.

Ainsi, le vecteur associé à un lexème est généré dynamiquement par le modèle en fonction de la phrase entière. Cette approche permet de lever l'ambiguïté lexicale et de capturer des relations syntaxiques et sémantiques plus fines. La différence majeure avec les plongements lexicaux statique réside donc dans le fait que `BERT` encode le sens contextuel d'un lexème, permettant d'obtenir plusieurs représentations d'un même lexème selon le contexte.

Les plongements contextuels s'éloignent de la définition de la ressource lexicale classique, dans la mesure où l'accès à l'information linguistique passe par l'utilisation d'un modèle. En ce sens, ils constituent davantage un outil de génération d'informations qu'une ressource lexicale statique.

1.6 La place des ressources lexicales aujourd'hui en TAL

Si l'évolution des représentations distributionnelles, de `word2vec` (Mikolov et al., 2013) aux modèles contextuels `BERT` (Devlin et al., 2019), puis aux architectures génératives de type `GPT` (Brown et al., 2020) suggère une capacité croissante à capturer les connaissances sémantiques de manière implicite, elle remet en question la place des ressources lexicales dans l'ère neuronale.

Dans cette section, nous nous intéressons à la place des ressources lexicales en TAL. Nous décrivons dans un premier temps, l'intégration des ressources lexicales à différents types de grands modèles de langue. Dans un second temps, nous discutons plus spécifiquement des enjeux pour les langues peu dotées, dont le traitement est plus difficile en raison du manque de données. Enfin, nous ouvrons la discussion au devenir des ressources lexicales dans le contexte des grands modèles de langue, omniprésents dans le paysage du TAL actuel.

1.6.1 Intégration de ressources lexicales aux grands modèles de langue

Les approches fondées sur des ressources lexicales présentent l'avantage d'une explicitation des connaissances, là où les grands modèles de langue reposent sur des approximations statistiques. Des travaux se sont alors intéressés à des méthodes hybrides qui combinent connaissances explicites et apprentissage à partir de corpus.

Un premier ensemble de travaux s'est concentré sur l'intégration des ressources lexicales au niveau des plongements lexicaux statiques. Une méthode connue est le *retrofitting* (Faruqui et al., 2015) qui consiste à ajuster *a posteriori* des vecteurs `word2vec` afin de rapprocher les représentations de mots liés, tels que définis dans `WordNet` et `FrameNet`. L'avantage de cette

méthode est qu'elle permet d'améliorer les représentations distributionnelles générées à partir de connaissances explicites, sans avoir à ré-entraîner le modèle.

Yu and Dredze (2014) proposent une autre approche permettant d'apprendre des plongements lexicaux de `word2vec` en incluant des ressources sémantiques telles que `WordNet` et `Paraphrase Database` (Ganitkevitch et al., 2013) dans la fonction objectif. Dans cette approche, les ressources lexicales ne sont pas utilisées pour l'évaluation mais comme élément central du processus d'apprentissage des représentations.

Avec l'apparition des grands modèles de langue contextuels, l'accent a été mis sur le traitement de la polysémie et des informations sous-lexicales. Si les modèles type `BERT` capture le contexte, ils présentent des difficultés pour traiter la polysémie et les entités de domaines spécifiques.

Levine et al. (2020) ont abordé cette question avec `SenseBERT`. Ils démontrent que le masquage `BERT` standard ne parvient pas à saisir les catégories sémantiques de haut niveau. Leur contribution consiste à modifier l'objectif de pré-entraînement afin que le modèle prédise non seulement le mot masqué, mais aussi son *supersense* à partir de `WordNet`, permettant ainsi des améliorations sur la tâche de désambiguisation lexicale et de *word-in-context*. Dans la même veine, Aloui et al. (2020) combinent les *supersenses* et les plongements contextuels avec un modèle hybride appelé `SLICE`, pour répondre au problème d'interprétabilité. Leur méthode faiblement supervisée permet d'apprendre des représentations interprétables à partir de corpus bruts et de petites listes de mots en français. Les ressources utilisées sont le `WOLF` (Sagot and Fišer, 2008), un `WordNet` du français et `FrSemCor` (Barque et al., 2020), un corpus annoté du français en *supersenses*.

Étant entraînés sur des textes non structurés et non étiquetés, les modèles de langue contextuels ne contiennent aucune référence explicite aux entités du monde réel et sont souvent incapables de mémoriser des faits concernant ces entités (Logan et al., 2019). Pour répondre à cette problématique, Peters et al. (2019) créént `KnowBERT`, constitué d'un modèle `BERT` dans lequel sont intégrés `WordNet` et `Wikipedia`. `KnowBERT` démontre une amélioration des performances sur des tâches en aval telles que l'extraction de relations, le typage d'entités et la désambiguïsation lexicale.

Enfin, dans le contexte des grands modèles de langue auto-régressifs, l'intégration des ressources lexicales tend à se faire de manière plus indirecte, notamment *via* des stratégies de *prompting* ou de récupération d'information externe (Lewis et al., 2020). Plutôt que d'être intégrées au cœur du modèle, les ressources lexicales peuvent être mobilisées comme sources de connaissances explicites lors de l'inférence, par exemple sous forme de définitions (Lu et al., 2024), de paraphrases (Wahle et al., 2024) ou de relations lexicales (Kanayama et al., 2024) injectées dans les *prompts*.

L'intégration de ressources peut également avoir lieu dans la phase d'affinage, comme la version affinée de `LLaMA-2-7b`, `TaxoLLaMA` (Moskvoretskii et al., 2024). Le modèle est affiné à partir de données de `WordNet`. Les échantillons d'entraînement se concentrent sur les paires hyponymes-hyperonymes et comprennent des définitions afin d'aider le modèle à lever l'ambiguïté des sens des mots pendant l'entraînement. Le modèle a été évalué à travers quatre tâches sémantiques lexicales principales : découverte d'hyperonymes, enrichissement taxonomique, implication lexicale et construction taxonomique et montrent des résultats état-de-l'art sur 11 des 16 configurations expérimentales.

1.6.2 Enjeux pour les langues peu dotées

La question des langues peu dotées constitue un enjeu majeur pour l'évaluation de la place des ressources lexicales dans le TAL contemporain. Bien que les grands modèles de langue soient souvent présentés comme multilingues, leurs performances varient fortement selon les langues, en raison d'un déséquilibre important dans les données d'entraînement (Joshi et al., 2020). Les langues disposant de faibles ressources textuelles sont généralement sous-représentées, ce qui limite la capacité des modèles à apprendre des régularités lexicales et morphosyntaxiques. Dans ce contexte, les ressources lexicales peuvent fournir des connaissances structurées difficilement déductibles à partir de corpus réduits.

Plusieurs travaux ont montré que l'intégration de lexiques ou de ressources sémantiques permet d'améliorer les performances pour des langues peu dotées, notamment dans des tâches de désambiguïsation lexicale, de traduction automatique ou d'analyse morphologique (Navigli and Ponzetto, 2012). Des ressources multilingues comme `BabelNet`, qui combine `WordNet` et `Wikipedia`, illustrent le potentiel de ces approches pour réduire les écarts de performance entre les langues.

Dans le cas des grands modèles de langues auto-régressifs plus particulièrement, Zhang et al. (2024) proposent `LingoLLM`, une approche exploitant des dictionnaires bilingues, des grammaires descriptives et des analyseurs morphologiques pour permettre à `GPT-4` et `Mixtral-8x7B` pré-entraînés de traiter des langues inconnues durant l'entraînement [20]. `LingoLLM` repose sur l'exploitation des capacités d'apprentissage en contexte (*in-context learning*) des grands modèles de langue *instruction-tuned* et ne nécessite pas d'entraînement.

Sakajo et al. (2025) utilisent également des dictionnaires bilingues avec une approche de transfert de vocabulaire qui aligne de manière itérative un sous-mot dans une langue cible à des sous-mots dans une langue source et initialise les plongements des sous-mots cibles.

Les travaux sur les langues peu-dotées insistent sur la nécessité d'approches hybrides combinant les modèles de langue et des ressources lexicales dans la mesure où celles-ci peuvent constituer des points d'ancrage linguistiques explicites, facilitant l'adaptation ou le contrôle du comportement du modèle pour des langues faiblement représentées.

1.6.3 Les grands modèles de langue ont-ils des « compétences lexicales » ?

Les performances élevées des grands modèles de langue auto-régressifs sur un large éventail de tâches soulèvent la question de leur « compétence lexicale » et, plus largement, de leur statut vis-à-vis des connaissances linguistiques.

Plusieurs auteurs distinguent les capacités langagières, entendues comme l'aptitude à produire des énoncés plausibles, des capacités linguistiques au sens strict, qui impliquent une connaissance explicite et systématique des structures de la langue (Bender et al., 2021; Beguš et al., 2025). Cette distinction est théorisée sous les termes de « compétence linguistique formelle » (connaissance des règles et motifs statistiques) *versus* « compétence linguistique fonctionnelle » (utilisation du langage dans le monde réel), les grands modèles de langue démontrant des facilités dans la première mais présentant des lacunes notables dans la seconde (Mahowald et al., 2024). Dans cette perspective, les grands modèles de langue apparaissent comme des systèmes statistiques puissants, capables de reproduire des régularités observées dans les données (Bender et al., 2021),

20. Les huit langues considérées sont le manchu, le gitksan, l'arapaho, le bribri, le tsez, le wolof, l'uspanteko et le natugu.

mais dont la maîtrise de distinctions linguistiques reste variable selon la tâche, et surtout selon la langue.

Les ressources lexicales, quant à elles, reposent généralement sur des cadres théoriques et visent à fournir une description systématique et interprétable de la langue. La montée en puissance des grands modèles de langue a conduit certains travaux à considérer ces modèles eux-mêmes comme des ressources linguistiques, voire comme des formes implicites de théories linguistiques émergentes (Manning et al., 2020; Ambridge and Blything, 2024).

Cette évolution s'accompagne toutefois de risques, notamment la marginalisation de formes rares, qui sont moins bien représentées dans les données d'entraînement et donc moins susceptibles d'être produites par les modèles. Guo et al. (2025) montrent en effet que les grands modèles de langue tendent à homogénéiser la production langagière en privilégiant les tournures fréquentes au détriment de la diversité linguistique et des *outliers* créatifs. À long terme, cette dynamique soulève des questions sur l'impact des systèmes neuronaux sur la diversité linguistique et sur la place des descriptions linguistiques explicites dans le TAL.

Enfin, le relatif désengagement vis-à-vis des ressources lexicales peut s'expliquer par un changement de paradigme : là où les ressources visaient initialement à décrire la langue et ses différents niveaux d'analyse, les grands modèles de langue privilégient une approche purement distributionnelle, fondée sur la fréquence et la cooccurrence des structures. Si cette approche permet d'atteindre des performances élevées sur de nombreuses tâches, et sur des langues très dotées, elle ne garantit pas une couverture équilibrée des phénomènes linguistiques. Dans ce contexte, les ressources lexicales conservent un rôle essentiel, non seulement comme outils d'enrichissement des modèles, mais aussi comme références théoriques pour l'évaluation des modèles (Ciaccio et al., 2025).

1.7 Conclusion

Les ressources lexicales sont des objets essentiels à la description de la langue. Dans ce chapitre, nous avons décrit l'évolution des ressources lexicales sur deux domaines étroitement liés, la linguistique et le TAL, passant d'un statut d'empreintes de théories linguistiques à celui de matériau fondamental pour les systèmes automatiques de TAL. Ces ressources peuvent prendre différentes formes, allant de dictionnaires à des réseaux sémantiques, et décrivent des informations relatives à divers niveaux linguistiques telles que la morphologie, la syntaxe ou la sémantique.

L'essor des modèles de langue neuronaux a marqué un tournant majeur dans l'utilisation des ressources lexicales en TAL, en transformant les modes de représentation de l'information linguistique. Celle-ci n'est plus encodée sous forme symbolique, directement lisible par l'humain, mais sous forme de représentations numériques destinées aux machines, soulevant de nouveaux enjeux en termes d'interprétabilité et de reproductibilité.

Le chapitre aborde également la question de la place des ressources lexicales aujourd'hui en TAL, depuis l'émergence des grands modèles de langue qui semblent capables d'inférer les informations linguistiques à partir de grandes quantités de données. Nous notons toutefois que la qualité et la représentativité de ces données sont questionnables, en raison des architectures opaques de ces systèmes et que ces approches ne sont aujourd'hui pas applicables à toutes les langues, en particulier les langues peu dotées.

Chapitre 2

25 ans de ressources lexicales du français à TALN

Sommaire

2.1 Introduction 37
2.2 Des ressources lexicales du français libres d'utilisation 38
2.2.1 Ressources lexicales classiques 38
2.2.2 Plongements lexicaux statiques 39
2.3 Le corpus TALN 2001-2025 40
2.3.1 Constitution du corpus 40
2.3.2 Pré-traitements et annotations manuelles 41
2.4 Évolution diachronique des ressources 42
2.5 Évolution des applications 45
2.5.1 Construction de ressources spécialisées 45
2.5.2 Construction d'outils 45
2.5.3 Correction d'erreurs et évaluation 46
2.5.4 Des ressources symboliques aux plongements lexicaux statiques . . 47
2.6 La question de la réutilisabilité 47
2.7 Conclusion 49

2.1 Introduction

Dans le chapitre 1, nous avons présenté une description historique des ressources lexicales à travers la linguistique et le traitement automatique des langues (TAL). Dans ce chapitre, nous nous proposons d'observer plus spécifiquement, l'évolution diachronique des ressources lexicales du français. Pour ce faire, nous nous inspirons des travaux de Mariani et al. (2019b) qui consistent en une analyse bibliométrique à partir du corpus `NLP4NLP`, comprenant les articles publiés dans 34 conférences et revues du domaine du TAL sur une période de 50 ans. Par le biais de ce corpus, les auteurs fournissent diverses statistiques sur le nombre d'articles et d'auteurs au fil du temps et montrent l'évolution des thèmes de recherche, l'identification des auteurs qui les ont introduits et l'utilisation des ressources langagières.

Dans le même esprit, nous nous intéressons à la présence des ressources lexicales du français sur une période de 25 ans dans la conférence nationale de TAL, TALN. La conférence

TALN est née en 1994 dans le cadre du pôle Langage Naturel du GdR-PRC Communication Homme-Machine [21] et a lieu annuellement depuis 1997. Depuis 1999, elle se déroule sous l'égide de l'ATALA, l'association Association pour le Traitement Automatique des Langues [22].

À partir d'un corpus d'articles scientifiques extraits de TALN entre 2000 et 2025, nous cherchons à répondre aux questions suivantes :

— Quelles sont les ressources du français les plus utilisées en TAL ? Pour quelles applications ?

— Quelles causes sont à l'origine de la création de nouvelles ressources ?

— Quels sont les critères influençant la réutilisabilité d'une ressource ?

Dans la section 2.2, nous présentons la liste des ressources lexicales du français que nous retenons pour notre analyse. Nous nous concentrons sur les ressources libres d'utilisation et nous distinguons les ressources lexicales symboliques, des ressources neuronales (plongements lexicaux statiques). Dans la section 2.3, nous détaillons la construction du corpus TALN 2001-2025, constitué de 1 687 articles et le processus d'annotations manuelles que nous avons conduit pour extraire les articles présentant des ressources lexicales.

Nous observons ensuite l'évolution diachronique des ressources, quantitativement parlant, en section 2.4 puis celle des applications en section 2.5. Enfin, nous abordons la question de la réutilisabilité à la section 2.6.

La méthodologie et une partie des expériences de ce chapitre reprennent celles présentées dans Choi et al. (2023).

2.2 Des ressources lexicales du français libres d'utilisation

2.2.1 Ressources lexicales classiques

Établir une liste exhaustive des ressources lexicales du français constitue une tâche délicate dans la mesure où un certain nombre de ressources existent mais sont peu connues, voire inaccessibles en terme de droit d'utilisation. De ce fait, nous considérons ici uniquement les ressources lexicales du français libres d'utilisation disponibles sur la plate-forme d'outils et de ressources linguistiques de la langue française ORTOLANG [23] et la LRE Map [24], plate-forme internationale de suivi de l'utilisation et de la création des ressources langagières. Nous présentons ces deux plateformes ci-après.

ORTOLANG pour Outils et Ressources pour une Traitement Optimisé de la LANGue est une infrastructure hébergeant des données (corpus, lexiques, dictionnaires etc.) et des outils sur la langue française dans le but de valoriser et partager les données sur le français et les langues de France, ainsi que de favoriser leur réutilisation par les laboratoires publics et les partenaires industriels.

Au moment de la sélection des ressources sur la plateforme en 2023, nous avons décidé de ne pas prendre en compte les lexiques spécialisés (pour l'orthophonie ou la phonologie articulatoire), ni les lexiques de langues régionales (lexiques du patois de Valbonnais ou de Saint-Martin-la-Porte).

21. `https://www.atala.org/-Conference-TALN-RECITAL`
22. `https://www.atala.org/association`
23. `https://www.ortolang.fr/market/lexicons`
24. `https://lremap.elra.info/?type=Lexicon&availability=Freely+Available&languages=french`

LRE Map est une initiative de ELRA (European Language Resources Association), l'association européenne des ressources langagières dont la mission principale est de mettre à la disposition de la communauté des ressources langagières pour les technologies des langues (*Human Language Technologies*). La LRE Map est ainsi décrite comme un mécanisme destiné à contrôler l'utilisation et la création de ressources langagières en recueillant des informations sur les ressources existantes et nouvellement créées. Pour compléter notre liste, nous avons considéré les ressources lexicales librement disponibles du français dont les liens étaient accessibles en 2023.

Vingt-six ressources lexicales du français. Au total, 26 ressources disponibles sur ORTOLANG et la LRE Map sont recensées pour notre analyse (cf. Tableau 2.1).

Domaine linguistique	Type	Ressource
Morphologie	Dictionnaire	`DELA`
	Lexique	`Morphalou` (ATILF, 2019)
		`VerNom` (Missud et al., 2020)
		`Lexique4linguists` (Schalchli, 2022)
	Réseau	`Démonette` (Hathout and Namer, 2014)
Morphologie/Syntaxe	Dictionnaire	`Dictionnaire morphosyntaxique du français (DM)` (Trouilleux, 2012)
	Lexique	`Le`*`fff`* (Clément et al., 2004; Sagot, 2010)
		`Marsalex` (Blache and Rauzy, 2008)
		`GLÀFF` (Sajous et al., 2013)
		`FLELex` (François et al., 2014)
Syntaxe	Dictionnaire	`Dicovalence` (Van den Eynde and Mertens, 2006, 2010; Mertens, 2010)
	Lexique	`Lexique-Grammaire` (Gross, 1975)
		`VfrLPL` (Rauzy and Blache, 2007)
		`LGLex` (Constant and Tolone, 2010)
Syntaxe/Sémantique	Lexique	`LVF` (Dubois and Dubois-Charlier, 1997)
		`Nomage` (Balvet et al., 2011)
		`Verb∋Net` (Danlos et al., 2014)
	Réseau	`Holinet` (Prost, 2022)
Sémantique	Dictionnaire	`Dictionnaire Électronique des Synonymes (DES)` (Ploux and Victorri, 1998)
		`Prolexbase` (Tran and Maurel, 2006)
	Lexique	`ReSyf` (Gala et al., 2013)
		`French FrameNet` (Candito et al., 2014)
	Réseau	`rezoJDM` (Lafourcade and Joubert, 2008; Lafourcade and Le Brun, 2020)
		`WOLF` (Sagot and Fišer, 2008)
		`RL-fr` (Lux-Pogodalla and Polguère, 2011)
		`DBnary` (Sérasset, 2015)

TABLE 2.1 – Liste des 26 ressources lexicales traditionnelles regroupées par domaine et par type.

2.2.2 Plongements lexicaux statiques

Grâce à leurs représentations vectorielles codant des informations linguistiques, les plongements lexicaux statiques s'apparentent à une nouvelle forme de ressource lexicale, rendue célèbre avec les modèles `word2vec` (Mikolov et al., 2013), puis `fastText` (Bojanowski et al., 2017) et `GloVe` (Pennington et al., 2014). Comme mentionné dans la section 1.5 du chapitre 1, les plongements lexicaux contextuels s'éloignent de la définition de ressource lexicale en raison de son caractère dynamique obtenu à partir d'un modèle, contrairement à des plongements lexicaux statiques qui peuvent être récupérés tels quels comme une définition ou une information d'un

lexique. Nous ne les considérons donc pas dans notre analyse.

Les plongements lexicaux statiques sont obtenus à partir de réseaux de neurones à deux couches qui ont pour but de représenter le contexte linguistique d'un lexème dans une représentation vectorielle. Pour cela, un modèle est entraîné sur un grand corpus de textes et génère une représentation vectorielle de chaque lexème, i.e. un plongement lexical.

Si les premiers plongements lexicaux statiques ont été créés pour l'anglais, cette approche a été rapidement étendue à d'autres langues dont le français. Nous observons ainsi les occurrences des plongements lexicaux statiques du français `word2vec` (Fauconnier, 2015) et `fastText` (Grave et al., 2018).

2.3 Le corpus TALN 2001-2025

Notre étude étant focalisée sur les ressources lexicales du français, nous décidons de constituer un corpus d'articles publiés à TALN, la conférence nationale française de TAL.

Par ailleurs, l'extraction d'articles présentant des ressources nécessitant une phase d'annotations manuelles, notre méthodologie est difficilement adaptable à une conférence plus grande telle que LREC (International Conference on Language Resources and Evaluation) [25], la conférence internationale sur les ressources langagières et leur évaluation, où le nombre d'articles est bien plus conséquent.

2.3.1 Constitution du corpus

Pour constituer le corpus, nous récupérons automatiquement les fichiers PDF des actes de TALN publiés sur ACL Anthology [26] de 2001 à 2025. Les fichiers sont ensuite convertis au format XML avec la librairie GROBID (Lopez, 2023).

De 2001 à 2025, un total de 2 619 articles a été publié, en comptant les actes de TALN, RECITAL, les ateliers et les éditions conjointes avec CORIA et JEP. La conversion avec GROBID pouvant être défectueuse, une vérification semi-automatique est faite en deux temps : nous déterminons si la balise *lang* correspond à celle du français ou de l'anglais et si des mots-outils de la langue concernée sont présents dans le corps de l'article, ce qui nous permet de vérifier si la conversion a bien produit du texte.

Nous recensons ainsi 80 articles mal convertis dont 39 pour l'année 2013 et 24 articles pour l'année 2001 [27].

Par ailleurs, nous faisons le choix de conserver uniquement les articles publiés à TALN et à RECITAL, ce qui exclut les articles JEP, CORIA et des ateliers, ceux-ci pouvant biaiser les résultats. En effet, en 2014, deux ateliers sur les ressources ont eu lieu (FondamenTAL et RLTLN), qui ont largement participé à augmenter les chiffres obtenus. À partir de 2024, les *abstracts* traduits d'articles publiés à des conférences internationales sont également acceptés. Les *abstracts* étant généralement très courts, ils ne nous permettent pas de juger de la réelle utilisation d'une ressource dans les travaux. Nous décidons donc de ne pas les prendre en compte.

25. `https://www.elra.info/elra-events/lrec/`
26. `https://aclanthology.org/venues/jeptalnrecital/`
27. Nous notons que le corpus ACL Anthology (Rohatgi, 2022) présente les mêmes articles défectueux. L'erreur de conversion provient vraisemblablement d'un problème dans les articles d'origine.

Le corpus final [28] contient **1 687 articles** exploitables au format XML. La figure 2.1 présente le nombre d'articles du corpus TALN 2001-2025 par année.

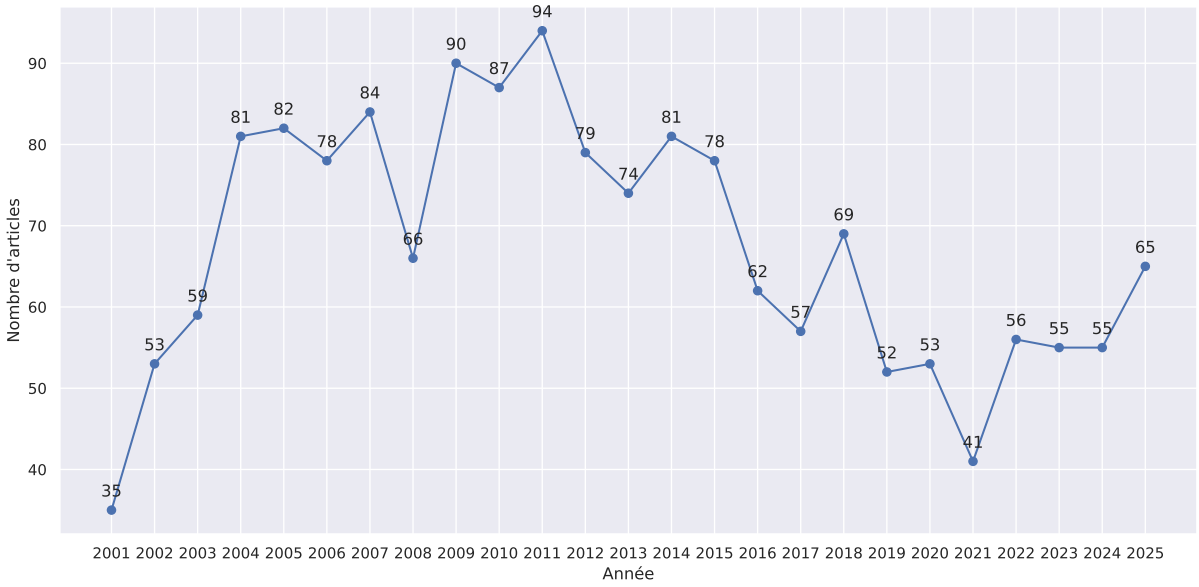


FIGURE 2.1 – Nombre d'articles du corpus TALN 2001-2025 par année.

2.3.2 Pré-traitements et annotations manuelles

Afin d'observer l'utilisation des ressources sélectionnées dans les articles, les occurrences de chaque ressource dans le corps de l'article sont extraites automatiquement. Nous décidons de considérer uniquement le corps de l'article avec la balise *body* pour éviter les biais dus aux occurrences dans les résumés et la bibliographie. En amont, des pré-traitements simples tels qu'une suppression de certaines ponctuations, des majuscules et une tokenisation sont appliqués. Nous veillons également à normaliser autant que faire se peut les noms des ressources. Par exemple, *glaff* est remplacé par *glàff*, plus fréquent.

Au total, seuls **246 articles** présentent au moins une des ressources considérées, ce qui représente un peu moins de 15 % du corpus entier.

Les occurrences brutes ainsi extraites ne nous permettant pas de tirer une quelconque conclusion, nous décidons d'effectuer une phase d'annotations manuelles sur les articles présentant au moins une occurrence d'une ressource. Pour chaque article et chaque ressource présente dans celui-ci, nous attribuons une classe parmi les cinq suivantes :

— **Construction-Extension** : l'article présente comment la ressource a été élaborée ou traite de son extension ou amélioration.

— **Utilisation** : l'article présente une expérience où la ressource est utilisée pour une tâche spécifique (construction d'une autre ressource, étiquetage morpho-syntaxique...).

— **Comparaison** : l'article utilise la ressource comme une référence pour faire une comparaison avec une autre ressource.

— **Mention** : l'article ne fait que mentionner la ressource (état de l'art).

28. Le corpus est librement disponible sur le dépôt suivant : `https://gitlab.inria.fr/papers2/taln2023`.

— **Erreur** : l'article présente une occurrence incorrecte (à cause d'une erreur dans le fichier XML), l'occurrence ne correspond pas au nom de la ressource (par exemple, « wolf » peut désigner un nom propre), la ressource ne concerne pas le français (par exemple, `FrameNet` de l'anglais).

L'annotation est effectuée en séparant les ressources lexicales non neuronales des plongements lexicaux statiques.

Pour les ressources lexicales symboliques, un accord inter-annotateurs est calculé sur un échantillon de 20 articles avec deux paires d'annotateurs (annotateur 0 - annotateur 1, annotateur 0 - annotateur 2), résultant en un accord parfait sur 19 articles. En retirant les articles qui contiennent des erreurs, nous obtenons un total de **166 articles** présentant une ressource traditionnelle.

Quant aux plongements lexicaux, nous veillons à faire la distinction entre la ressource et l'outil. En effet, nous annotons en « Construction » le cas où les auteurs génèrent des plongements lexicaux et les mettent à disposition, contrairement au cas où le modèle est utilisé pour une application précise sans redistribution des plongements générés. Par ailleurs, comme pour les ressources multilingues, nous ne prenons en compte que les articles traitant du français. Nous précisons également que cette vérification ne s'est pas faite de manière triviale en raison d'un certain nombre d'articles ne mentionnant pas explicitement la langue traitée, un problème récurrent bien identifié en TAL (Ducel et al., 2022).

Les articles contenant des plongements lexicaux sont annotés par trois annotateurs. Un accord inter-annotateurs a été calculé de la même manière que précédemment et montre un accord pour 17 articles sur 20 articles. Au total, nous comptons **33 articles** présentant des plongements lexicaux statiques du français (contre 65 articles pour l'anglais).

2.4 Évolution diachronique des ressources

D'un point de vue diachronique, le nombre d'articles présentant des ressources symboliques augmente progressivement de 2001 à 2014, où un pic est atteint avec 27 articles (cf. Figure 2.2).

De 2014 à 2020, le nombre d'articles diminue mais se maintient autour d'une dizaine d'articles puis diminue drastiquement en 2021, 2022 et 2023. Nous remarquons toutefois une légère hausse de l'utilisation des ressources lexicales symboliques en 2024 et 2025.

Concernant la construction de ressources, les articles sur le sujet ont été publiés entre 2007 et 2020, notamment en 2014 où plusieurs ressources voient le jour comme le `French FrameNet` (Candito et al., 2014) et `Demonette` (Hathout and Namer, 2014). Cette dernière étant une fusion de ressources existantes, le nombre de ressources utilisées augmente également. D'autres ressources font également l'objet d'extension comme le `RL-fr` (Lux-Pogodalla and Polguère, 2011) ou `RezoJDM` (Lafourcade et al., 2014).

Cette dynamique est représentative des changements d'approches de la recherche en TAL au fil du temps. Les ressources lexicales ont constitué des objets indispensables pour l'élaboration de systèmes à base de règles ou hybrides dans les années 2000 à 2015, expliquant ainsi la tendance à la création de nouvelles ressources. À partir de 2017, un nouveau type de ressource lexicale apparaît avec les plongements lexicaux statiques (cf. Figure 2.3). Si ces derniers sont moins utilisés que les ressources lexicales symboliques entre 2017 et 2020, ils se font davantage présents dans

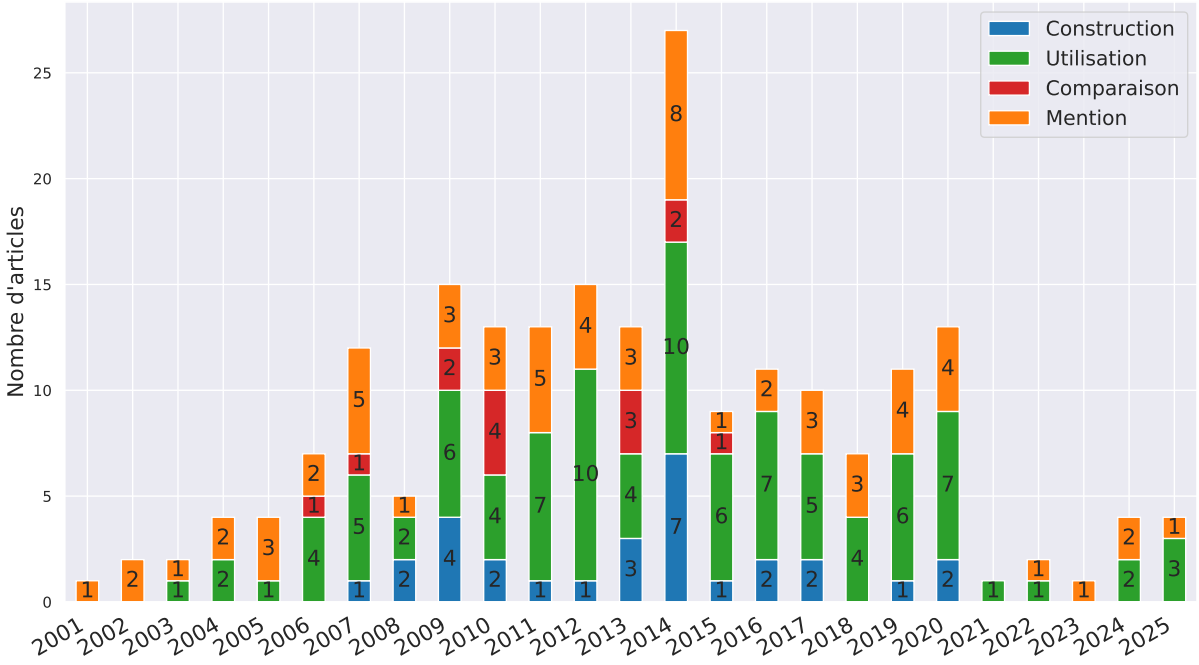


FIGURE 2.2 – Distribution dans le temps du nombre d'articles présentant une ressource lexicale traditionnelle selon le type d'usage.

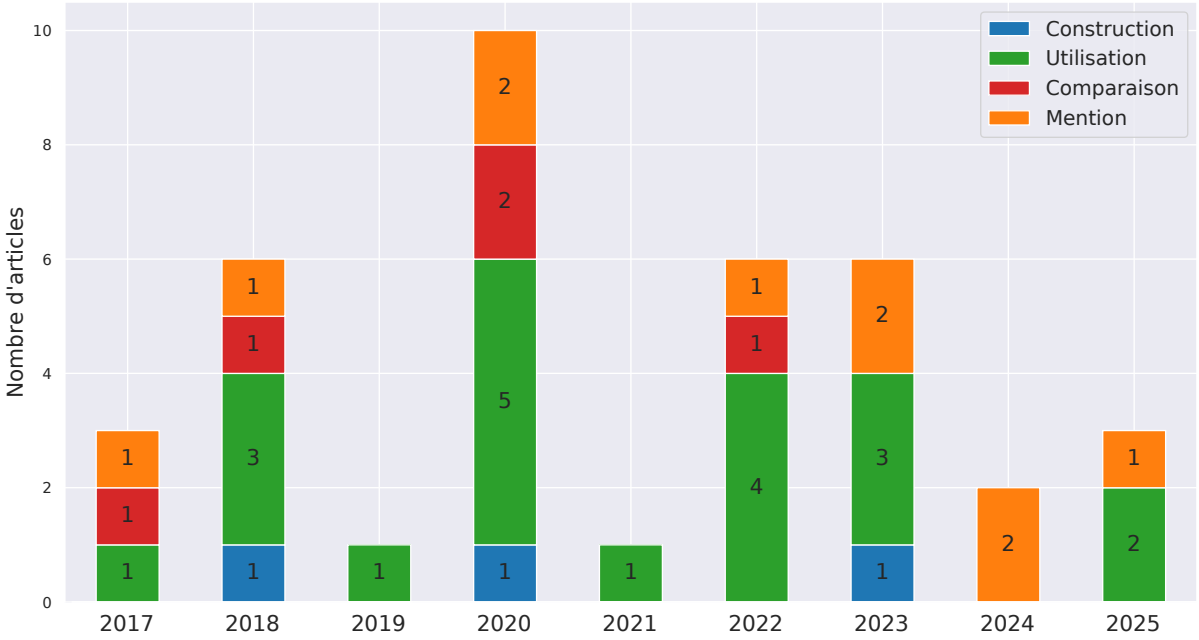


FIGURE 2.3 – Distribution dans le temps du nombre d'articles présentant des plongements lexicaux selon le type d'usage.

les années 2022 et 2023 avec 12 articles au total contre trois pour les ressources symboliques ces années-là.

Nous nous intéressons ensuite aux ressources les plus présentes dans notre corpus décrites dans les figures 2.4 et 2.5.

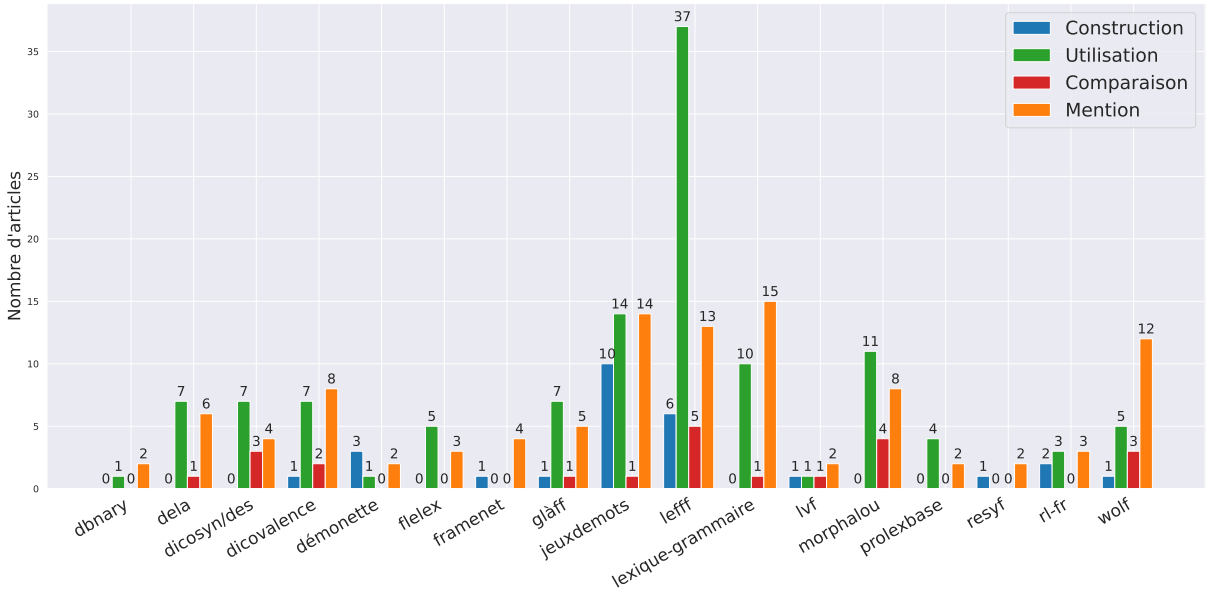


FIGURE 2.4 – Distribution des ressources symboliques les plus fréquentes selon le type d'usage.

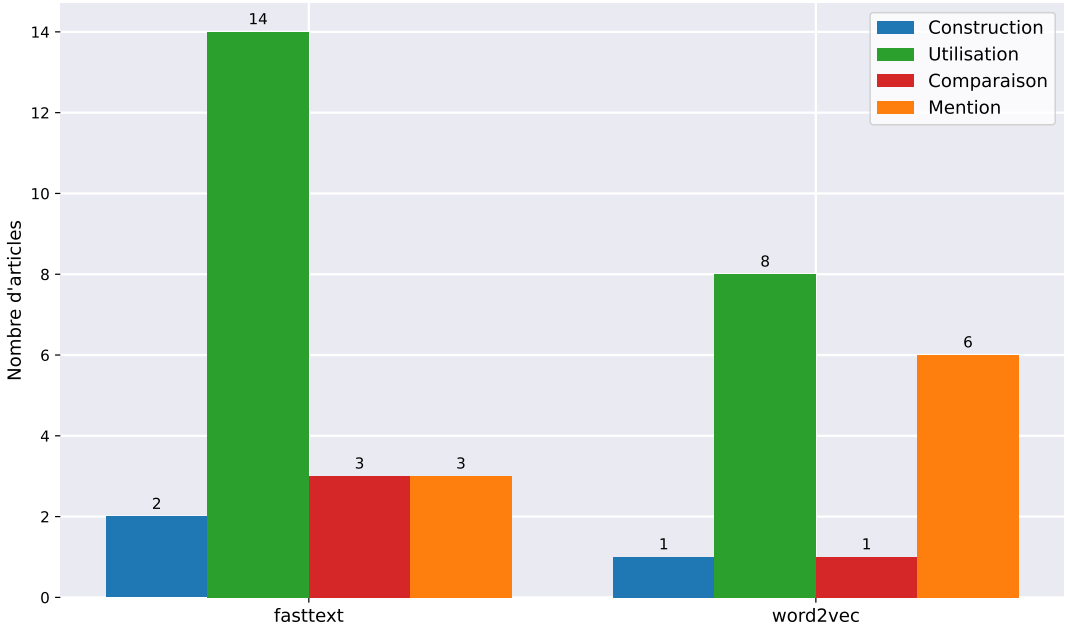


FIGURE 2.5 – Distribution des plongements lexicaux selon le type d'usage.

La figure 2.4 présente la distribution des ressources lexicales symboliques selon le type d'usage telle qu'identifiée manuellement [29]. Les ressources les plus présentes dans le corpus sont le `Lefff`

29. Par souci de lisibilité, nous ne présentons pas les neuf ressources restantes, leurs occurrences étant strictement inférieures à 2, toutes classes confondues.

avec 61 articles, JeuxDeMots (39 articles), le Lexique-Grammaire (26 articles), Morphalou (23 articles) et WOLF (21 articles), toutes mentionnées une quinzaine de fois. La tendance est toutefois différente quant à leur utilisation, le Le*fff* se démarquant largement avec 37 articles.

Concernant les plongements lexicaux statiques, la Figure 2.5 montre que les plongements fastText créés par Grave et al. (2018) ont été davantage utilisés que ceux de word2vec.

2.5 Évolution des applications

Dans cette section, nous présentons des exemples d'application des ressources de notre inventaire. Par ces exemples, nous cherchons à lister les applications les plus fréquentes et à observer leur évolution avec l'apparition de nouvelles ressources telles que les plongements lexicaux mais également les nouvelles approches basées sur les modèles de langue.

2.5.1 Construction de ressources spécialisées

Avant l'ère du TAL neuronal basé sur l'apprentissage de grands corpus de texte, les ressources lexicales ont permis de construire des systèmes symboliques basés sur les règles. Avoir à disposition des ressources décrivant la langue de manière adaptée pour une tâche est alors une problématique majeure.

En morphosyntaxe, les ressources telles que les lexiques morphosyntaxiques ont donc largement été exploités pour la création de nouvelles ressources spécialisées.

Le Le*fff*, ressource la plus utilisée dans les articles de TALN, a permis de créer un **lexique de l'ancien français** (Sagot, 2019), un **lexique du français préclassique** (Diwersy et al., 2015) et un **lexique d'adjectifs dénominaux** (Strnadová and Sagot, 2011). Ces deux derniers lexiques ajoutent des entrées de Morphalou en complément. Les informations syntaxiques du Le*fff* et de Dicovalence ont également servi à la construction de PrepLex, un **lexique de prépositions** pour l'analyse syntaxique (Fort and Guillaume, 2007).

Plus récemment, le Lexique-Grammaire a été utilisé pour la création de PolylexFLE, une **base de données polylexicales** pour le FLE (Français Langue Etrangère) (Todirascu et al., 2019). En 2020, Étienne et al. (2020) ont extrait les formes fléchies du Le*fff* pour enrichir un **lexique d'émotions** pour l'annotation d'un corpus.

En sémantique, Bigeard (2017) a utilisé le réseau lexico-sémantique RezoJDM pour construire un **lexique spécialisé de termes et expressions désignant une maladie ou un trouble de l'humeur** pour la détection des mésusages de médicaments fréquents.

2.5.2 Construction d'outils

Au-delà de créer de nouvelles ressources, les ressources lexicales ont également permis d'élaborer des systèmes automatiques traitant des tâches spécifiques de TAL.

Une tâche traditionnelle en morphosyntaxe est l'étiquetage morpho-syntaxique consistant à associer aux lexèmes d'un texte les informations grammaticales correspondantes comme son étiquette de partie du discours, son genre et son nombre. Pour automatiser cette tâche, divers **étiqueteurs morpho-syntaxiques** du français ont été créés, notamment à partir du Le*fff*,

tels que MELT$_{fr}$ (Denis and Sagot, 2010), Macaon (Nasr et al., 2011), Talismane (Urieli and Tanguy, 2013) et SEM (Tellier et al., 2014). L'étiqueteur LGTagger inclut lui aussi le Le*fff* en plus d'autres ressources comme DELA et Prolex (Constant and Sigogne, 2011).

La tâche de lemmatisation a également entraîné la création de **lemmatiseurs**, permettant de récupérer les lemmes à partir de dictionnaires et lexiques morphologiques tels que DELA, Le*fff* et Morphalou (Panchenko et al., 2013; Tanguy et al., 2019).

Par ailleurs, des **outils pour l'extraction et l'annotation d'informations linguistiques** spécifiques ont profité des ressources. On peut citer par exemple un outil d'annotation automatique pour les pronoms anaphoriques et impersonnels à partir du Lexique-Grammaire (Danlos, 2005), un outil de détection des cadres valenciels à partir de Dicovalence (Nasr and Béchet, 2009), un outil d'extraction d'expressions polylexicales grâce à FleLEX (Tack et al., 2016) et enfin dernièrement, un outil de détection de genre à partir du Le*fff* et du GLÀFF (Flesch and Burnett, 2025).

Des outils pour des tâches sémantiques ont également été mis au point. Ramadier and Lafourcade (2018) ont exploité les relations entre des termes génériques et des termes médicaux spécifiques de RezoJDM pour élaborer un **système de simplification lexicale** pour des comptes-rendus radiologiques. Sur une toute autre tâche, Plu et al. (2018) proposent JeuxDeLiens, un **système de désambiguïsation d'entités nommées** dans des documents en français.

Enfin, Gala et al. (2014) s'appuient sur diverses ressources lexicales pour construire un **modèle de prédiction de complexité lexicale**. Des ressources comme Morphalou sont alors utilisées d'une part pour constituer le corpus d'apprentissage du modèle et d'autre part pour extraire des variables caractérisant la complexité lexicale. Les informations morphologiques sont notamment extraites de Morphalou et Polymots (Gala and Rey, 2008). Quant aux variables sémantiques, elles proviennent des réseaux lexicaux RezoJDM et BabelNet.

2.5.3 Correction d'erreurs et évaluation

Les ressources lexicales constituent des sources de connaissances linguistiques fiables permettant la correction de sorties générées automatiquement. Ainsi, Tanguy et al. (2025) s'appuient sur Morphalou pour vérifier et corriger la lemmatisation générée par Stanza (Qi et al., 2020).

En outre, les ressources lexicales sont particulièrement importantes pour l'évaluation des systèmes automatiques. Elles sont ainsi fréquemment utilisées pour la constitution d'échantillons de test pour des tâches variées.

En 2011, Henestroza Anguiano and Denis (2011) utilisent le WOLF et l'EuroWordNet pour l'évaluation d'un thésaurus distributionnel construit automatiquement. En traduction automatique, le Le*fff* a été exploité pour l'évaluation morphologique (Burlot and Yvon, 2018).

Gosset et al. (2021) élaborent un système d'extraction automatique de relations sémantiques, évalué sur des relations d'hyperonymie et d'hyponymie de RezoJDM.

Enfin, plus récemment, Petrov et al. (2025) créent ALF, un jeu de données analogiques à partir du RL-fr pour l'évaluation de la connaissance lexicale des grands modèles de langue.

2.5.4 Des ressources symboliques aux plongements lexicaux statiques

Avec l'apparition des modèles `word2vec` (Mikolov et al., 2013), le TAL a vu un nouveau type d'approche émerger, passant des approches par apprentissage supervisé basées sur les ressources symboliques à des approches neuronales basées sur l'apprentissage principalement non supervisé. Appliquer ces modèles nécessite alors de représenter les lexèmes sous forme de vecteurs, appelés plongements lexicaux.

Ainsi, à partir de 2017, nous pouvons observer un certain nombre de travaux exploitant des modèles neuronaux tels que des BiLSTM intégrant en entrée des plongements lexicaux. En effet, les plongements lexicaux statiques pré-entraînés, tels que `word2vec` et `fastText`, peuvent être utilisés directement, évitant l'entraînement d'une couche de plongements dans le modèle. Dans notre corpus TALN 2001-2025, les travaux utilisant ces modèles traitent de différentes tâches comme la détection du sexisme (Chiril et al., 2019), l'identification d'erreurs d'annotations (Randriatsitohaina and Hamon, 2020), la reconnaissance d'entités nommées (Dupont, 2017; Gabay et al., 2022) et la classification de questions (Eshkol-Taravella et al., 2022).

Le développement de l'architecture transformeur et des grands modèles de langues change le paysage du TAL. Contrairement aux réseaux de neurones traditionnels qui traitent les données de manière séquentielle, ces modèles reposent uniquement sur des mécanismes d'attention qui permettent la parallélisation, améliorant ainsi l'entraînement (Vaswani et al., 2017).

Les modèles basés sur les plongements lexicaux statiques sont alors davantage utilisés pour constituer des *baselines*, comme le font Petrov et al. (2025) dans la tâche de résolution d'analogies et (Cao et al., 2023) pour la reconnaissance d'entités nommées.

Si le fait d'encoder des informations linguistiques pour un lexème leur confère le statut de ressource lexicale, les plongements lexicaux statiques pré-entraînés sont relativement peu réutilisés tels quels. En effet, les modèles sont davantage utilisés en tant qu'outil pour générer des plongements lexicaux propres à une application.

2.6 La question de la réutilisabilité

Les résultats présentés dans les sections précédentes montrent que certaines ressources lexicales sont davantage utilisées que d'autres, comme par exemple le `Lefff`. Se pose alors la question de la réutilisabilité et les raisons qui pousseraient à réutiliser une ressource plutôt qu'une autre.

La notion de réutilisabilité peut être vue sous deux angles (Ooi, 1998) :

— **Réutilisation de ressources existantes :** une ressource est réutilisée pour une autre application que celle pour laquelle elle a été initialement créée.

— **Multifonctionnalité :** une ressource est créée de manière à ce qu'elle soit réutilisée dans diverses applications.

Un exemple notable d'une ressource réutilisable selon la première définition qui en est donnée est `WordNet`. Initialement créé par des psycholinguistes pour étudier les théories de la mémoire sémantique, il est aujourd'hui l'une de ressources les plus utilisées en traitement automatique des langues (Mariani et al., 2019a,b), notamment grâce à la richesse des informations linguistiques

disponibles (concept de *synset* illustré par des définitions et des relations sémantiques), à une licence libre et à son intégration dans des outils de TAL tels que la librairie NLTK [30].

Pour le français, il est difficile de déterminer une ressource de référence en TAL malgré les efforts de la communauté pour créer de nouvelles ressources. Plusieurs raisons peuvent expliquer ce constat :

— une couverture trop faible,

— des informations linguistiques trop spécifiques,

— une ressource peu accessible : problèmes de licence, de mise à jour, de format,

— une ressource peu adaptable à d'autres applications.

On peut ainsi supposer que l'absence d'une ressource générale de référence pousse les chercheurs à créer des ressources correspondant à leurs besoins, sans nécessairement optimiser leur réutilisabilité.

Pour vérifier cette hypothèse, nous nous intéressons aux quatre ressources les plus utilisées : Le*fff*, RezoJDM, Morphalou et le Lexique-Grammaire. Nous cherchons à observer si les utilisations d'une ressource sont menées par des personnes impliquées dans la construction de celle-ci. Pour déterminer la liste des créateurs d'une ressource, nous récupérons les auteurs des articles de référence. Dans le cas où la ressource ne fait pas l'objet d'une publication, nous prenons en compte les personnes présentées comme contributrices dans les sites officiels des ressources [31]. Le tableau 2.2 présente ces quatre ressources et les personnes impliquées dans leur création.

Ressource	**Personnes impliquées dans la création**
Le*fff*	Benoît Sagot, Lionel Clément
RezoJDM	Mathieu Lafourcade, Nathalie Le Brun, Alain Joubert
Morphalou	Sandrine Ollinger, Christophe Benzitoun, Marie Tonnelier, Evelyne Jacquey, Ulrike Fleury
Lexique-Grammaire	Maurice Gross et al. (LADL) [32]

TABLE 2.2 – Créateurs des ressources les plus utilisées dans le corpus TALN 2001-2025.

D'après le tableau 2.3, le Le*fff*, ressource la plus utilisée dans notre corpus, est majoritairement utilisée par des personnes extérieures à sa création avec 27 utilisation externes contre 10 internes. Le Le*fff* a la caractéristique d'être une des premières ressources créées pour le TAL avec une approche semi-automatique permettant une large couverture.
De même, Morphalou présente davantage d'utilisations externes qu'internes.

Inversement, RezoJDM est la seule ressource dont le nombre d'utilisations internes est supérieur au nombre d'utilisations externes (8 contre 6). Ces huit articles comptent Mathieu Lafourcade en auteur, principal fondateur de la ressource. RezoJDM a fait l'objet d'un certain nombre de travaux, se démarquant par sa richesse et sa densité considérable. Créé par myriadisation et par des mécanismes semi-automatiques, le réseau permet notamment le développement de travaux en syntaxe et en sémantique. Toutefois, sa richesse peut être à la fois une qualité comme un

30. https://www.nltk.org/howto/wordnet.html

31. Nous précisons que les informations sur les concepteurs des ressources peuvent contenir des erreurs.

32. Les personnes considérées sont des personnes passées par le LADL ayant utilisé la ressource comme L. Danlos, E. Tolone, S. Voyatzi, T. Nakaruma, E. Laporte, S. Paumier. Cette liste n'est pas exhaustive.

Ressource	Utilisation externe	Utilisation interne	Total
Le*fff*	27	10	37
RezoJDM	6	8	14
Morphalou	10	1	11
Lexique-Grammaire	5	5	10

TABLE 2.3 – Nombre d'utilisations internes ou externes pour les quatre ressources les plus utilisées dans le corpus TALN 2001-2025.

défaut dans la mesure où une ressource de ce calibre (plusieurs millions de nœuds et de relations) s'avère difficile à manipuler.

2.7 Conclusion

Grâce aux efforts de la communauté française de TAL, de nombreuses ressources librement disponibles ont été créées pour le français. Dans ce chapitre, nous avons observé leur évolution dans la recherche en TAL sur une période de 25 ans. Pour cela, nous avons constitué un corpus d'articles scientifiques extraits et annotés manuellement, nous permettant d'analyser l'évolution de leur présence dans les articles et de leur application.

Les résultats montrent notamment que la présence des ressources a été croissante jusqu'en 2014, puis a diminué à partir de 2015. Cette tendance peut être corrélée avec l'expansion des modèles de langue neuronaux, qui exploitent davantage de grandes quantités de textes que des ressources lexicales. Par ailleurs, nous notons que le Le*fff* se démarque des autres ressources avec un fort nombre d'articles l'utilisant. Nous pouvons attribuer ce résultat à la grande couverture de la ressource, tant en termes de taille que de contenu linguistique, qui a permis la construction de ressources spécialisées et d'outils morphosyntaxiques.

L'utilisation des ressources lexicales a changé parallèlement à l'évolution des techniques de TAL. Si elles ont permis à créer initialement des ressources et des outils tels que des étiqueteurs morphosyntaxiques ou des lemmatiseurs, elles sont aujourd'hui davantage exploitées comme référence pour l'évaluation des systèmes automatiques.

Deuxième partie

Enrichir des réseaux lexico-sémantiques par prédiction de liens

Introduction

La création de ressources lexicales de manière manuelle par des linguistes constitue un travail chronophage et coûteux. La couverture de ces ressources devient alors une problématique majeure dont l'automatisation peut permettre une expansion plus rapide.

Dans cette partie, nous nous intéressons à la complétion de réseaux lexico-sémantiques du français *via* des modèles de prédictions de liens. Les réseaux lexico-sémantiques et plus généralement les graphes de connaissances sont connus pour être incomplets dans la mesure où atteindre l'exhaustivité dans la description du monde ou d'une langue est impossible. Pour pallier cette incomplétude, plusieurs tâches ont été mises au point dans le domaine de l'affinage des graphes de connaissances (*Knowledge Graph Refinement*), dont la tâche de prédiction de liens (Paulheim, 2017).

Nous consacrons ainsi le chapitre 3 à un état de l'art de la complétion des graphes de connaissances et des réseaux lexico-sémantiques à travers différentes approches et plus particulièrement la prédiction de liens. Nous présentons le cadre de la tâche avec les principaux jeux de données et systèmes conçus, ainsi que les métriques d'évaluation.

Bien que l'incomplétude des graphes de connaissances et des réseaux lexico-sémantiques soit un argument avancé dans la majorité des travaux sur la prédiction de liens, nous constatons que très peu aboutissent à un réel enrichissement des ressources. En effet, la configuration générale adoptée consiste à récupérer une partie du graphe de connaissances pour constituer l'échantillon sur lequel est évalué un système de prédiction de liens. Si le système est effectivement évalué à inférer des informations inconnues, la ressource n'est toutefois pas augmentée par de nouvelles instances. Contrairement à l'assertion initiale sur l'incomplétude des graphes, on retrouve un objectif davantage tourné vers la construction et l'évaluation de modèles que la complétion de la ressource. Nous cherchons alors à explorer si un modèle de prédiction de liens peut prédire de potentiels nouveaux triplets pour un réseau lexico-sémantique du français (QR 3).

Dans le chapitre 4, nous proposons de traiter la tâche de prédiction de liens avec une approche orientée vers la ressource et non vers les modèles. Notre but est de regarder au-delà des performances et d'analyser les prédictions des systèmes pour extraire de potentielles nouvelles informations pour enrichir un réseau lexico-sémantique parcimonieux du français, le `RL-fr`. Pour ce faire, nous évaluons dans un premier temps sept modèles de prédiction de liens sur deux réseaux lexico-sémantiques de densité différente : le `RL-fr` et `RezoJDM`. Dans un second temps, nous générons des prédictions robustes du modèle le plus performant avec la technique de Monte-Carlo Dropout (Gal and Ghahramani, 2016), robustesse que nous évaluons à travers des analyses manuelles.

Chapitre 3

État de l'art : complétion de graphes par prédiction de liens

Sommaire

3.1 Introduction 55
3.2 Graphes de connaissances et réseaux lexico-sémantiques 56
3.2.1 Définitions et notations 56
3.2.2 Contexte, construction et applications 58
3.3 Pallier l'incomplétude des graphes de connaissances 61
3.3.1 Complétion externe 61
3.3.2 Complétion interne 62
3.4 Prédiction de liens 63
3.4.1 Description de la tâche 63
3.4.2 Jeux de données 64
3.4.3 Métriques d'évaluation 65
3.4.4 Méthodes de prédiction de liens 66
3.5 Conclusion 70

3.1 Introduction

La représentation d'informations sous forme de graphe permet de modéliser les relations entre les entités décrites, qu'elles soient des unités lexicales dans les réseaux lexico-sémantiques ou des informations du monde réel dans les graphes de connaissances.

Dans le domaine du traitement automatique des langues (TAL), la construction et l'utilisation de réseaux lexico-sémantiques se sont largement développées, notamment grâce à `WordNet` (Miller, 1995), une ressource créée manuellement par des psycholinguistes, mais qui est aujourd'hui l'une des plus utilisées en TAL (Mariani et al., 2019a). Outre les connaissances linguistiques, des travaux en TAL ont exploité les graphes de connaissances qui représentent des informations du monde réel ou de domaine spécifique dans un format lisible par une machine (Schneider et al., 2022).

Un des enjeux principaux dans la construction des graphes de connaissances et des réseaux lexico-sémantiques consiste à obtenir une couverture importante, difficile à atteindre à travers un

travail manuel qui s'avère coûteux et chronophage. Des travaux se sont donc intéressés à combiner l'extraction de données provenant de sources diverses *via* des techniques (semi-)automatiques pour construire des graphes à grande échelle (Paulheim, 2017).

Cependant, ces ressources sont connues pour être incomplètes dans la mesure où atteindre l'exhaustivité dans la description du monde ou d'une langue est impossible. Au-delà de la problématique de l'incomplétude, les graphes de connaissances et les réseaux lexico-sémantiques peuvent également nécessiter des corrections d'erreurs générées lors de leur construction, ainsi qu'une mise à jour des informations, afin d'éviter toute obsolescence.

Dans cette thèse, nous aborderons uniquement la problématique de l'incomplétude des graphes. Nous consacrons ainsi le chapitre 3 à un état de l'art sur la complétion des graphes, et plus particulièrement la complétion à travers la tâche de prédiction de liens.

Pour ce faire, nous commençons par définir dans la section 3.2 les concepts de graphe de connaissances et de réseaux lexico-sémantiques, la distinction étant peu marquée dans la littérature. Nous présentons également un point de vue historique sur la construction de ces objets et leurs applications actuelles en TAL. Ensuite, la section 3.3 se focalise sur la problématique de l'incomplétude des graphes et présente une vue d'ensemble des approches de complétion. Enfin, nous détaillons plus spécifiquement la tâche de prédiction de liens dans la section 3.4, qui inclut un état de l'art sur les jeux de données, les métriques d'évaluation et les méthodes existantes.

3.2 Graphes de connaissances et réseaux lexico-sémantiques

3.2.1 Définitions et notations

Pour des raisons de clarté, nous proposons ici de définir les concepts et de poser les notations que nous utiliserons tout au long de cette thèse.

Graphe. En théorie des graphes, un graphe est défini comme une structure (V,E) composée d'un ensemble de nœuds V (V pour *vertice*) et d'un ensemble d'arêtes $E \subseteq V \times V$ (E pour *edge*) (Bonfante et al., 2018). Les nœuds représentent des entités et les arêtes, les relations entre les entités. Un graphe peut être orienté ou non orienté.

Réseau. La théorie des réseaux est une sous-partie de la théorie des graphes. Un réseau est donc défini comme un graphe dont les nœuds et/ou les arêtes possèdent des attributs [33]. Les attributs peuvent représenter des informations comme des types ou des poids.

En TAL, la distinction entre un réseau et un graphe s'estompe avec les graphes de connaissances, ces derniers pouvant également présenter des attributs de nœuds et d'arêtes. Les deux termes sont souvent retrouvés de manière interchangeable.

Dans cette thèse, nous ne faisons pas de réelle distinction entre les concepts de graphe et de réseau, nous ferons davantage la différence entre deux objets : les graphes de connaissances et les réseaux lexico-sémantiques (cf. Figure 3.1).

Graphe de connaissances. Bien que le terme « graphe de connaissance », *Knowledge Graph* (KG) en anglais, soit aujourd'hui largement utilisé, sa définition formelle reste toutefois litigieuse (Hogan et al., 2021). Nous utilisons la définition large proposée par Hogan et al. (2021)

33. `https://en.wikipedia.org/wiki/Network_theory`, consulté le 25 août 2025.

qui définit un graphe de connaissance comme un « graphe de données destiné à accumuler et à transmettre des connaissances sur le monde réel, dont les nœuds représentent des entités d'intérêt et dont les arêtes représentent les relations entre ces entités. ».

Les connaissances représentées dans ces graphes vont de faits simples tels que « Albert Einstein est né dans l'Empire allemand » à des faits quantifiés comme « tous les chiens sont des mammifères ».

Les graphes de connaissances permettent de fournir une représentation structurée de connaissances factuelles utiles pour des systèmes de TAL comme les systèmes de recommandations ou de questions-réponses.

Réseau lexico-sémantique. Techniquement, un réseau lexico-sémantique peut être considéré comme un sous-type de graphe de connaissances [34]. Plus précisément, un réseau lexico-sémantique décrit des connaissances linguistiques : les nœuds représentent des unités lexicales et les arêtes encodent des relations sémantiques entre les unités lexicales telles que la synonymie, l'antonymie ou l'hyperonymie. Les informations décrites sont par exemple « voiture est synonyme d'automobile » ou « roue est un méronyme de voiture ».

Ainsi, les systèmes tirant parti des réseaux lexico-sémantiques traitent davantage des thèmes liés à la langue comme la désambiguïsation lexicale ou l'inférence lexicale.

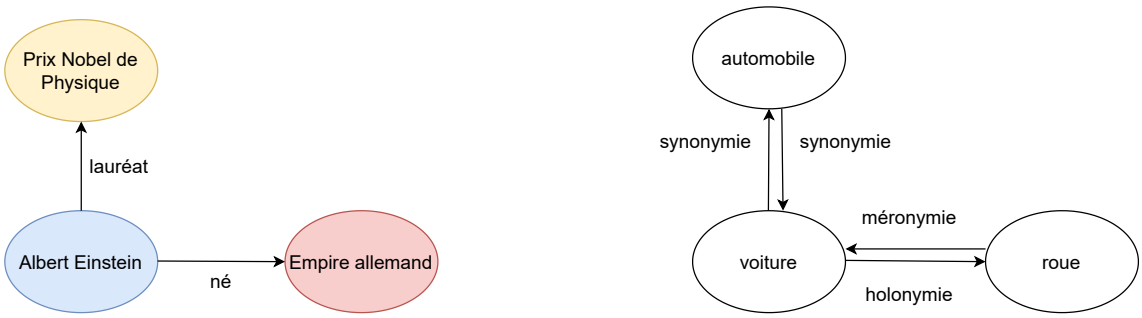


FIGURE 3.1 – Exemple d'un graphe de connaissance (gauche) et d'un réseau lexico-sémantique (droite).

Formellement, un graphe de connaissance KG est composé d'un ensemble de nœuds V (*Vertices*) et un ensemble d'arêtes E (*Edges*), noté $KG = (V, E)$.

Il peut également être décrit par un ensemble de triplets $T = (h, r, t)$; $h \in V, t \in V, r \in R$ ainsi qu'une fonction $\phi : (h, t) \in V \times V \rightarrow r \in R$.

h et t désignent respectivement ce que nous appellerons l'entité-source et l'entité-cible du triplet. La relation entre ces deux entités est désignée par r, et est donnée par ϕ.

Que ce soit dans un graphe de connaissances ou un réseau lexico-sémantique, un triplet encode un fait. Dans la figure 3.1, le graphe de connaissance présente trois entités de types différents et deux relations, liés en triplets :

— `(Albert Einstein, né, Empire allemand)`

— `(Albert Einstein, lauréat, Prix Nobel de Physique)`

34. `https://en.wikipedia.org/wiki/Semantic_network`, consulté le 25 août 2025.

Dans le réseau lexico-sémantique, les trois entités sont des unités lexicales liées par différentes relations sémantiques. Les triplets sont :

— `(voiture, synonymie, automobile)`

— `(automobile, synonymie, voiture)`

— `(roue, méronymie, voiture)`

— `(voiture, holonymie, roue)`

La figure 3.1 vise à illustrer en particulier les différences dans la nature des entités et de relations dans ce que nous considérons un graphe de connaissances du monde réel et un réseau lexico-sémantique. Ces graphes peuvent représenter des informations supplémentaires telles que des poids sur les arêtes, des entités de différents types (par exemple, des parties du discours pour un réseau lexico-sémantique) et des relations symétriques.

3.2.2 Contexte, construction et applications

Historiquement, le concept de graphes de connaissance a été introduit dans les années 1970 dans le cadre d'une discussion sur la manière de concevoir des systèmes pédagogiques modulaires (Schneider, 1973). À la fin des années 1980, les universités de Groningue et de Twente ont élaboré le projet `Knowledge Graphs` afin de représenter le contenu de textes médicaux et sociologiques sous forme de graphe. Le but était alors d'obtenir un système expert pour la recherche automatisée de relations causales ou pour l'aide à la décision (Sri Nurdiati and Hoede, 2008).

Les premiers graphes de connaissances sont ainsi davantage spécifiques à un domaine, avec des exemples notables comme `WordNet` (Miller, 1995; Fellbaum, 1998) pour la psycholinguistique ou Geonames[35] pour la géographie. Cette tendance est accentuée dans les années 2000 par un regain d'intérêt pour les ontologies, des représentations formalisées de connaissances spécifiques à un domaine utilisant des structures logiques.

Construction. Concevoir des ressources pour des connaissances générales a été l'objet de divers projets, notamment avec la création de bases de connaissances comme `Cyc` (Lenat, 1995), `SUMO` (Niles and Pease, 2001), `DBpedia` (Auer et al., 2007) et `Freebase` (Bollacker et al., 2007), mais ce n'est qu'à l'initiative de Google en 2012 que le concept de graphe de connaissance prend réellement forme avec le projet `Google Knowlege Graph`[36]. Le projet a pour but de créer automatiquement un graphe de connaissance à grande échelle en intégrant des milliards de faits provenant de différentes sources dont `Freebase` et `DBpedia`. La figure 3.2 présente un aperçu des graphes de connaissances populaires.

Concernant les réseaux lexico-sémantiques spécifiquement, l'utilisation massive de `WordNet` a entraîné la construction de `WordNets` dans d'autres langues (Bond and Paik, 2012), dont par exemple des langues européennes avec le projet `EuroWordNet` (Vossen, 1998) et les langues indiennes avec `IndoWordNet` (Bhattacharyya, 2010).

Outre ces nouveaux `WordNets`, `BabelNet` (Navigli and Ponzetto, 2012) a également été créé à partir de `WordNet`, ainsi que de `Wikipedia`, `OmegaWiki` et `Wiktionary`, constituant un vaste

35. `https://www.geonames.org/`

36. `https://blog.google/products/search/introducing-knowledge-graph-things-not/`, consulté le 26 août 2025.

Name	Instances	Facts	Types	Relations
DBpedia (English)	4,806,150	176,043,129	735	2,813
YAGO	4,595,906	25,946,870	488,469	77
Freebase	49,947,845	3,041,722,635	26,507	37,781
Wikidata	15,602,060	65,993,797	23,157	1,673
NELL	2,006,896	432,845	285	425
OpenCyc	118,499	2,413,894	45,153	18,526
Google's Knowledge Graph	570,000,000	18,000,000,000	1,500	35,000
Google's Knowledge Vault	45,000,000	271,000,000	1,100	4,469
Yahoo! Knowledge Graph	3,443,743	1,391,054,990	250	800

FIGURE 3.2 – Aperçu des graphes de connaissances populaires. Le tableau présente le nombre d'entités (*instances*), de triplets (*facts*), de types d'entités (*Types*) et de types de relations (*Relations*) (Tableau 1 de Paulheim (2017)).

réseau sémantique multilingue construit automatiquement à partir des ressources précédemment citées. Sa construction repose sur l'alignement automatique des *synsets* de `WordNet` avec des pages de `Wikipedia` et la fusion des concepts équivalents entre les langues (Navigli et al., 2021). Il en résulte un réseau unifié où chaque *Babel synset* relie des unités lexicales multilingues et des relations sémantiques de différents types (lexicales, sémantiques et encyclopédiques).

Un autre réseau sémantique multilingue notable est `SWOW` (De Deyne et al., 2019) (Small World of Words), réseau multilingue d'association de mots construit à partir d'expériences participatives à grande échelle. Les participants se voient présenter des mots-clés et sont invités à fournir les premiers mots qui leur viennent à l'esprit [37]. Les réponses sont agrégées et normalisées pour former un réseau orienté où le poids des liens reflète la fréquence des associations. Contrairement aux réseaux sémantiques classiques, `SWOW` cherche à capturer la structure cognitive du lexique mental et à modéliser les processus associatifs humains, pour des applications en psycholinguistique, sciences cognitives et sémantique distributionnelle.

Plus spécifiquement sur le français, nous pouvons citer deux réseaux lexico-sémantiques : le `RL-fr` (Lux-Pogodalla and Polguère, 2011) et `RezoJDM` (Lafourcade and Joubert, 2008). Le `RL-fr` (Réseau Lexical du Français) est un réseau lexical construit manuellement et basé sur la théorie Sens-Texte (Mel'čuk et al., 2012). Il est constitué de descriptions linguistiques qui rendent compte des propriétés combinatoires et syntaxiques des unités lexicales par l'encodage de fonctions lexicales.

`RezoJDM` est un autre réseau lexico-sémantique du français, construit *via* des jeux ayant un but (GWAP) sur la plate-forme `JeuxDeMots` et des mécanismes semi-automatiques. Les joueurs participent à des jeux en ligne qui collectent implicitement des associations lexicales, telles que des synonymes, des hyperonymes ou des liens thématiques. Cette méthode permet une collecte de données à grande échelle et une mise à jour en continu, sans coûts d'annotation manuelle directe (Lafourcade and Joubert, 2013; Lafourcade and Le Brun, 2020).

Le tableau 3.1 présente une vue d'ensemble de la composition des réseaux lexico-sémantiques décrits précédemment. S'ils partagent un même format de graphe, on observe une forte diversité dans les tailles et les informations représentées.

37. `https://smallworldofwords.org/en/project/research`.

Nom	**# Instances**	**# Triplets**	**# Type de relations**
`BabelNet` (v5.3)	22 892 310 *Babel synsets*	1 911 610 725	$\sim$ 200
`WordNet` (v3.0)	$\sim$ 117 000 *synsets*	978 636	26 $\sim$ 30
`RL-fr` (v3)	29 975 unités lexicales	76 464	692 fonctions lexicales
`RezoJDM`	$>$ 15 millions de termes	$>$ 1 milliard	$\sim$ 160
`SWOW-EN`	12 292 mots-clés	$\sim$ 3 684 600	1

TABLE 3.1 – Aperçu de la composition des réseaux lexico-sémantiques populaires.

Tâches et applications. D'après la taxonomie des tâches sur les graphes de connaissances de Schneider (1973) (cf. Figure 3.3), deux types de tâches se distinguent : celles pour l'acquisition de connaissances et celles pour leur application.

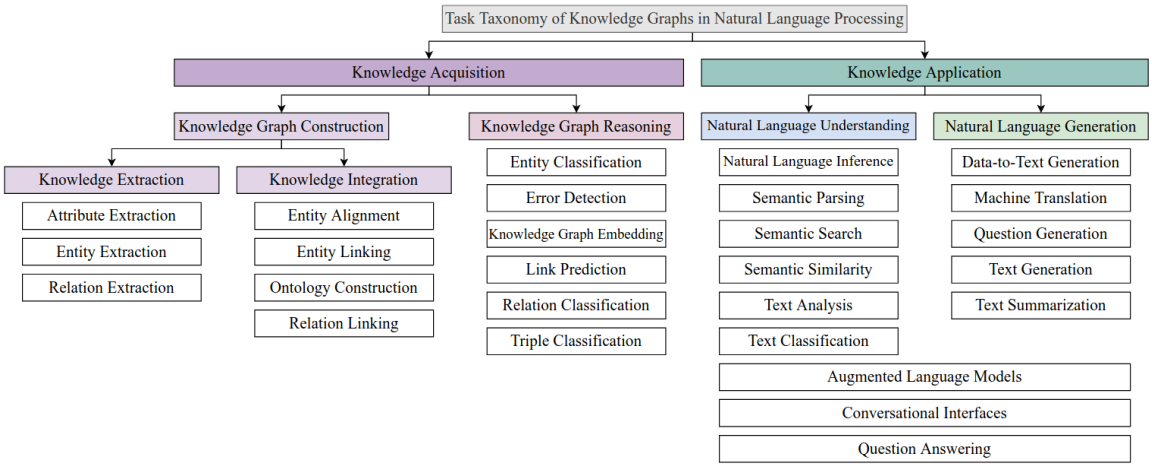


FIGURE 3.3 – Taxonomie des tâches impliquant les graphes de connaissances dans la littérature en TAL (Figure 1 de Schneider et al. (2022)).

L'acquisition de connaissances englobe des tâches cherchant à construire des graphes à partir de textes non structurés ou à conduire des raisonnements sur des graphes déjà construits. Les tâches de construction consistent en des tâches d'extraction (Al-Moslmi et al., 2020) ou d'intégration de connaissances (Chen et al., 2018; Zeng et al., 2021). Avec le développement de l'apprentissage profond et des réseaux de neurones, les tâches de raisonnement telles que la génération de plongements de graphe (Wang et al., 2017; Cai et al., 2018) et la prédiction de liens (cf. Section 3.4) se sont développées afin d'évaluer la cohérence et la représentation des informations d'un graphe.

Concernant les applications, les informations encodées dans les graphes de connaissances et les réseaux lexico-sémantiques sont exploitées pour la compréhension et la génération de la langue, dans des tâches en aval telles que les systèmes de question-réponse (Cimiano et al., 2013; Han, 2024), l'analyse et la classification de textes (Liu et al., 2023) ou la désambiguïsation lexicale (Kumar et al., 2019; Bevilacqua and Navigli, 2020). Enfin plus récemment, l'apparition des grands modèles de langue conduit à de nouvelles méthodes pour l'intégration des graphes de connaissances dans l'architecture et l'entraînement des modèles. On peut citer notamment

la génération à enrichissement contextuel, *Retrieval-augmented generation* (RAG) qui permet de récupérer dynamiquement des informations factuelles pertinentes dans le graphe de connaissances pour guider la génération des modèles (Lewis et al., 2020).

3.3 Pallier l'incomplétude des graphes de connaissances

Les graphes de connaissances et les réseaux lexico-sémantiques sont intrinsèquement incomplets pour diverses raisons, rendant impossible la description du monde ou de la langue de manière exhaustive. Parmi ces raisons, Hubert (2024) liste les plus courantes comme suit :

- **Des limites dans l'extraction automatique d'information** : les techniques d'extraction du TAL peuvent induire des erreurs et/ou des manques dans la complétion du graphe.
- **Une couverture sélective en fonction des intérêts et des priorités** : le graphe peut présenter une couverture disparate selon les thèmes en raison des choix et priorités des créateurs.
- **Des erreurs humaines dans la sélection des données** menant à des informations manquantes ou incorrectes.
- **Des informations dynamiques et évolutives** : les données évoluent dans le temps, particulièrement pour les connaissances du monde mais également pour le vocabulaire d'une langue. Les nouvelles informations doivent être modifiées et corrigées.

Des méthodes d'affinage (*Knowledge Graph Refinement*) ont été proposées pour traiter d'une part la complétion des graphes et d'autre part la correction d'erreurs (Paulheim, 2017). Dans cette thèse, nous traitons uniquement l'aspect complétion.

Le domaine de complétion des graphes de connaissances, *Knowledge Graph Completion* (KGC), s'intéresse à compléter le graphe de connaissances et à augmenter sa couverture en prédisant les relations potentielles entre les entités existantes et en découvrant de nouvelles entités et relations (Chen et al., 2020). Le but est d'augmenter la couverture du graphe pour des applications et des tâches en aval telles que les systèmes de recommandations, de questions-réponses, les réseaux sociaux en ligne, le e-commerce etc. Paulheim (2017) distingue deux types d'approches pour la complétion d'un graphe de connaissances :

- la complétion interne, qui consiste à s'appuyer uniquement sur le graphe et sa structure.
- la complétion externe, à l'inverse, compense les limites des informations internes au graphe en exploitant des ressources extérieures telles que des corpus de textes, des données du *web* ou d'autres graphes de connaissances.

3.3.1 Complétion externe

Les méthodes de complétion externe utilisent des informations provenant de ressources extérieures en les connectant au graphe de connaissances. Ces connexions peuvent prendre la forme de liens vers des pages *web*, telles que les pages Wikipedia reliées aux entités qu'elles décrivent ou d'interconnexions entre plusieurs graphes de connaissances (Paulheim, 2017).

Complétion à partir de données textuelles. Les premières approches pour enrichir un graphe de connaissances se basent sur des statistiques de cooccurrences dans des corpus de textes (Riloff and Shepherd, 1997; Biemann et al., 2004). Ces techniques nécessitent toutefois une intervention manuelle importante en raison de la faible qualité des prédictions.

Mintz et al. (2009) et Aprosio et al. (2013) utilisent la méthode de supervision à distance (*distant supervision*) pour enrichir respectivement `Freebase` et `DBpedia` à partir de corpus de textes provenant de `Wikipedia`. Dans un premier temps, l'approche consiste à relier les entités du graphe de connaissances à des corpus de texte *via* la reconnaissance d'entités nommées. Le but est ensuite de rechercher des motifs dans le texte correspondant à des types de relations. Par exemple, « le livre X de Y » peut être un motif de la relation « auteur » entre les entités X et Y dans le graphe. Ces motifs sont alors appliqués pour déterminer de nouvelles relations entre les entités du graphe.

Complétion à partir d'autres graphes de connaissances. Une approche majeure pour enrichir un graphe de connaissance consiste en l'alignement de graphes de connaissances, dans le but d'obtenir un graphe unifié. Les graphes de connaissances tels que `DBpedia`, `YAGO` ou `Freebase` contiennent des informations communes, voire complémentaires sur des entités équivalentes. Les premiers modèles d'alignement d'entités (*entity alignment*) vont des approches symboliques, fondées sur la correspondance sémantique et la similarité textuelle, aux approches neuronales ou basées sur des plongements (Suchanek et al., 2011; Zeng et al., 2021).

Exploiter un graphe de connaissances existant pour compléter un autre est également une approche adaptée dans un contexte multilingue, notamment dans la création de `WordNets` dans d'autres langues, à partir du `WordNet` anglais. Plusieurs `WordNets` ont ainsi été construits et/ou enrichis à travers des techniques d'alignement et de traduction automatique à partir de dictionnaires et corpus parallèles (Oliver and Climent, 2012; Apidianaki and Sagot, 2014; Lam et al., 2014; Arcan et al., 2016; Taghizadeh and Faili, 2016).

3.3.2 Complétion interne

Les méthodes de complétion interne s'appuient uniquement sur les informations présentes dans le graphe, comme sa structure ou les statistiques relatives aux faits déjà représentés. Les méthodes internes incluent :

- **les approches symboliques** fondées sur l'apprentissage de règles logiques ou de motifs relationnels, comme `AMIE` (Galárraga et al., 2013), `AMIE+` (Galárraga et al., 2015) ou `AnyBURL` (Meilicke et al., 2019). Ces systèmes apprennent des règles de Horn de la forme $r_1(x, y) \wedge r_2(y, z) \Rightarrow r_3(x, z)$ et les appliquent pour inférer de nouvelles arêtes.
- **les modèles basés sur les plongements de graphes**, qui représentent les entités et les relations dans un espace vectoriel continu afin de prédire de nouveaux liens plausibles.
- **les modèles neuronaux de graphes (GNN)** qui propagent l'information à travers les voisins pour capturer les dépendances contextuelles.

Ces deux dernières techniques se concentrent sur la tâche de prédiction de lien que nous détaillerons dans la section 3.4.

L'objectif commun des méthodes internes est de combler les lacunes du graphe initial en exploitant ses propriétés internes, de voisinage, les similarités de relations ou les motifs récur-

rents afin de générer de nouveaux triplets cohérents. Contrairement à la complétion externe, la complétion interne ne vise pas à augmenter la couverture du graphe à partir de connaissances nouvelles, mais à améliorer sa complétude intrinsèque et sa cohérence en exploitant sa structure.

3.4 Prédiction de liens

3.4.1 Description de la tâche

La prédiction de liens est une sous-tâche de la tâche générale de complétion de graphe. On distingue plus précisément trois sous-tâches classiques :

— La prédiction de liens (*Link Prediction*) vise à déterminer si un lien (relation) entre deux entités existe.

— La prédiction de relations (*Relation Prediction*) s'apparente à la prédiction de liens mais se focalise sur le type de relation à prédire : étant donné *(h, t)*, le modèle doit prédire *r*.

— La classification de triplets (*Triplet Classification*) est une tâche binaire consistant à prédire si un triplet est vrai ou faux.

Dans cette thèse, nous aborderons uniquement la tâche de prédiction de liens, et plus particulièrement sur la **prédiction de liens transductive**. En effet, il existe deux configurations pour la prédiction de liens : la prédiction transductive et la prédiction inductive.

En prédiction transductive, les entités prédites sont nécessairement des entités présentes dans le graphe, le modèle n'est pas en mesure de prédire de nouvelles entités. En pratique, l'ensemble des entités observées en phase d'apprentissage coïncide avec celles utilisées pour l'évaluation : le modèle apprend à compléter des relations manquantes entre des nœuds connus. Il se base alors sur des schémas de co-occurrences existants pour déduire les informations manquantes.

À l'inverse, la prédiction inductive cherche à généraliser le raisonnement à des entités ou sous-graphes non vus pendant l'entraînement (Galkin et al., 2022; Hubert et al., 2023). Le modèle doit donc apprendre des motifs relationnels ou structurels lui permettant d'inférer de nouveaux liens impliquant des nœuds inédits. Cette configuration reflète davantage les scénarios réels d'évolution des graphes de connaissances, où de nouvelles entités apparaissent continuellement. Les approches inductives, souvent fondées sur des réseaux de neurones de graphes ou des modèles de raisonnement local comme `GraIL` (Teru et al., 2020), mettent ainsi l'accent sur la capacité de généralisation plutôt que sur la simple mémorisation de co-occurrences observées comme c'est le cas en prédiction transductive.

En pratique, en prédiction transductive, étant donnés des triplets incomplets tels que `(h, r, ?)` ou `( ?, r, t)`, la tâche consiste à prédire l'entité manquante. Pour ce faire, les modèles de prédiction de liens sont entraînés à donner un meilleur score aux triplets positifs qu'aux triplets négatifs créés par *negative sampling* (Bordes et al., 2013), technique consistant à corrompre les triplets positifs en remplaçant l'entité-source ou l'entité-cible par une autre entité choisie au hasard.

À l'inférence, le modèle entraîné génère une distribution de probabilités sur toutes les entités dans le cas `(h, r, ?)`, puis classe les entités-cibles en fonction des probabilités (cf. Figure 3.4).

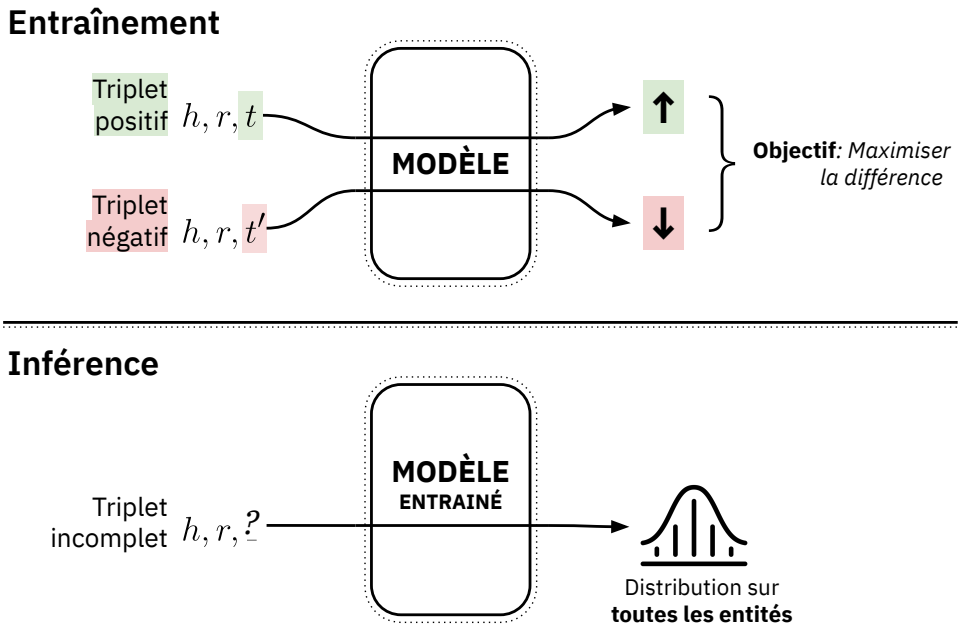


FIGURE 3.4 – Schéma des phases d'entraînement et d'inférence de la prédiction de liens.

3.4.2 Jeux de données

Il existe différents *benchmarks* pour l'évaluation des modèles de prédiction de liens selon le type de graphe comme par exemple, les réseaux de citations (`Cora`[38], `PudMed` (Xu et al., 2020)) ou les graphes spécifiques à des domaines (`BioKG` (Walsh et al., 2020) pour la biologie ou `EduKG` (Zhao et al., 2022) pour l'éducation).

En prédiction de liens transductive, les jeux de données les plus populaires proviennent de graphes de connaissances tels que `Freebase`, `WordNet`, `YAGO` et `NELL` (cf. Tableau 3.2).

Jeux de données	# entités	# triplets	# types de relation
`FB15k` (Bordes et al., 2013)	14 951	592 213	1 345
`FB15k-237` (Toutanova and Chen, 2015)	14 541	310 116	237
`YAGO3-10` (Mahdisoltani et al., 2013)	123 182	1 179 040	37
`WN18` (Bordes et al., 2013)	40 943	151 442	18
`WN18RR` (Dettmers et al., 2018)	40 943	86 835	11
`NELL-995` (Xiong et al., 2017)	75 492	154 213	200

TABLE 3.2 – Jeux de données pour la prédiction de liens.

FB15k et FB15k-237. Créé à partir de `Freebase`, `FB15k` contient les entités impliquées dans plus de 100 faits dans `Freebase` pour un total de 592 213 triplets (Bordes et al., 2013). Toutefois, Toutanova and Chen (2015) démontrent que le jeu de données présente une « fuite de données » (*test leakage*) puisqu'une partie des triplets de test sont déduits en inversant l'entité-source et

38. `https://graphsandnetworks.com/the-cora-dataset/`.

l'entité-cible de triplets de l'entraînement. Ils construisent alors `FB15k-237` pour y remédier, en retirant les relations inverses.

WN18 et WN18RR. Outre `FB15k` qui encode des informations du monde, Bordes et al. (2013) ont également créé `WN18`, un jeu de données lexicales à partir de `WordNet`, présentant 40 943 *synsets* et 151 442 triplets. Pour remédier au même problème de fuite de données, Dettmers et al. (2018) retirent toutes les relations inverses de `WN18`, donnant le nouveau jeu de données, `WN18RR`.

YAGO3-10. `YAGO3-10` (Mahdisoltani et al., 2013) est un jeu de données de connaissances générales (personnes, villes, pays, films et organisations) provenant de `YAGO3`, lui-même une extension du graphe de connaissance `YAGO` (Suchanek et al., 2007). Il contient des entités reliées par au moins dix relations différentes. Au total, `YAGO3-10` compte 123 182 entités, 37 types de relations et 1 179 040 triplets, dont la plupart décrit les attributs de personnes tels que la nationalité, le sexe et la profession.

NELL-995. `NELL` (*Never Ending Language Learning)* (Carlson et al., 2010; Mitchell et al., 2015) est un système fonctionnant en continu avec pour but d'extraire des informations factuelles à partir de pages web. Le jeu de données `NELL-995` est un sous-ensemble de la 995ème itération du système et compte 75 492 entités et 154 213.

3.4.3 Métriques d'évaluation

Inspirées du domaine de la recherche d'information (*Information Retrieval*), les métriques traditionnelles pour la tâche de prédiction de liens sont basées sur le classement des scores des prédictions correctes parmi toutes les prédictions générées.

Lors de la phase d'inférence, il est donné au modèle chaque triplet de test avec l'entité-cible masquée, `(h, r, ?)`. Étant donné ce triplet incomplet, le modèle attribue un score à toutes les entités-cibles possibles. Ces triplets sont ensuite classés selon leurs scores.

Mean Rank (MR) : étant donné un ensemble de triplets classés par leur score de prédiction, le MR calcule le rang moyen des triplets corrects. Un MR faible signifie une meilleure performance.

$$MR = \frac{1}{|Q|} \sum_{q \in Q} \text{rank}_q \qquad \in [1, |\mathcal{V}|]$$

Où Q est l'ensemble des triplets de test et rank_q est le rang de la bonne prédiction parmi les prédictions triées pour le triplet q. L'intervalle de cette métrique est $[1, |\mathcal{V}|]$ où $|\mathcal{V}|$ correspond au nombre d'entités dans le graphe.

Mean Reciprocal Rank (MRR) : MRR est la moyenne des inverses des rangs des triplets corrects. Elle permet d'obtenir une métrique dans l'intervalle $[0, 1]$ et rend possible la comparaison entre des jeux de données de différentes tailles. Un MRR proche de 1 indique une meilleure performance.

$$MRR = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{\text{rank}_q} \qquad \in [0, 1]$$

Hits@k : cette mesure calcule la proportion de triplets corrects apparaissant dans le top k de la liste classée des triplets. Les valeurs de k les plus courantes sont 1, 3 et 10, les scores augmentant logiquement avec la valeur de k. Une valeur *Hits@k* proche de 1 indique qu'une haute proportion des triplets sont correctement prédits.

$$Hits@k = \frac{1}{|Q|} \sum_{q \in Q} \mathbb{1}\left[\text{rank}_q \leq k\right] \qquad \in [0, 1]$$

Où $\mathbb{1}$ est la fonction caractéristique définie comme $\mathbb{1}\left[\text{rank}_q \leq k\right] = 1$ si q est classé entre les positions 1 et k, et 0 dans le cas contraire.

3.4.4 Méthodes de prédiction de liens

Le domaine de la prédiction de liens et plus généralement de la complétion de graphe a vu apparaître différents types de modèles avec l'évolution des modèles d'apprentissage automatique et plus récemment des grands modèles de langues.

Dans cette section, nous présentons les méthodes de complétion de graphe en les regroupant dans un premier temps en deux catégories définies dans l'étude de Chen et al. (2020) : les méthodes traditionnelles et les méthodes basées sur les plongements. Nous y ajoutons également une catégorie émergente depuis 2024, les méthodes basées sur les grands modèles de langues.

En raison du choix des modèles dans nos expériences (cf. Chapitre 4), nous apporterons davantage de détails aux méthodes basées sur les plongements.

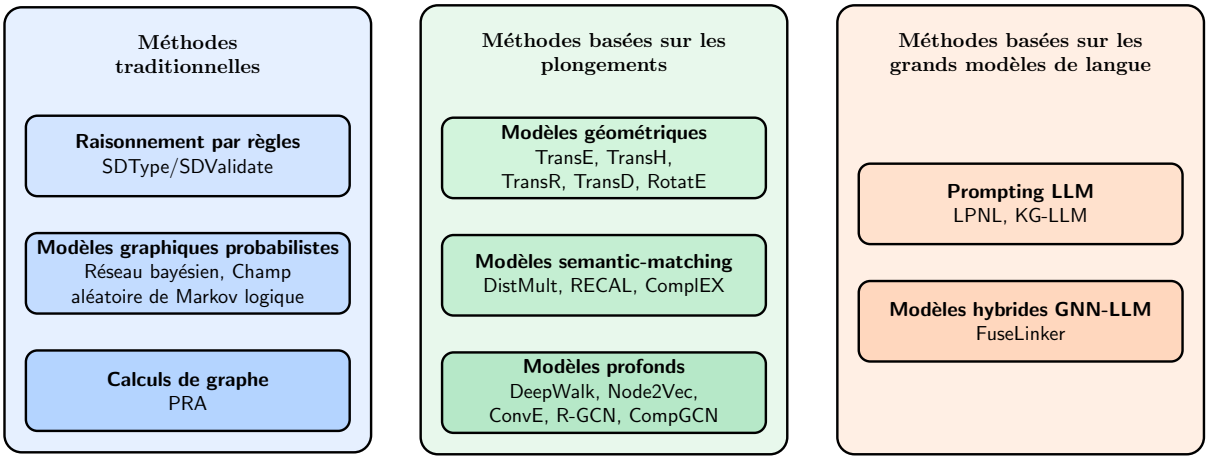


FIGURE 3.5 – Méthodes de complétion de graphe.

3.4.4.1 Les méthodes traditionnelles

Cette catégorie regroupe les méthodes suivantes :

— **les méthodes basées sur le raisonnement par règles :** ces méthodes visent à déterminer de nouveaux faits dans le graphe en utilisant des règles logiques et statistiques. Paulheim and Bizer (2014) proposent les algorithmes `SDType` et `SDValidate` (Statistic Distribution Type/Validate), qui consistent à prédire le type l'entité-source et/ou de l'entité-cible à partir d'une distribution statistique. Ces règles sont extraites manuellement ou à partir de la sémantique du graphe, permettant ainsi une forte interprétabilité. Toutefois, ces méthodes

ne sont pas applicables sur de grands graphes du fait des limites de calcul. Les méthodes basées sur le raisonnement délaissent de plus en plus les règles pour du raisonnement par représentations distributionnelles et réseaux de neurones.

— **les méthodes basées sur des modèles graphiques probabilistes :** un modèle graphique est une représentation en graphe d'objets probabilistes. Ces méthodes utilisent généralement des réseaux bayésiens (*Bayesian network*) et des champs aléatoires de Markov logiques (*Markov logic network*) permettant de combiner logique et probabilité pour des raisonnements incertains.

— **les méthodes basées sur des calculs de graphe :** ces méthodes se basent sur des caractéristiques statistiques des nœuds et des arêtes, telles que le degré entrant et/ou sortant des nœuds ou la matrice d'adjacence pour prédire de nouvelles entités ou relations. Un des algorithmes les plus célèbres est le *Path Ranking Algorithm* (PRA) (Gardner et al., 2013) qui utilise des marches aléatoires pour entraîner le modèle.

Si les méthodes traditionnelles ont l'avantage d'être interprétables, elles ne permettent pas de passer à l'échelle des graphes de connaissances de grande taille (Hubert, 2024).

3.4.4.2 Les méthodes basées sur les plongements

Ces méthodes, appelées également *embedding-based methods* ou *representation learning-based methods*, consistent à représenter les entités et les relations dans des espaces vectoriels de faible dimension. Elles ont l'avantage d'être applicables à des graphes de connaissances de grande taille et permettent de capturer des structures complexes et des relations sémantiques (Hubert, 2024).

Avec le développement de l'apprentissage machine et l'apparition de `word2vec` (Mikolov et al., 2013), ce type de modèles s'est largement développé à partir de 2013 et se divisent en trois sous-catégories :

— les modèles géométriques,

— les modèles *semantic-matching*,

— les modèles profonds.

Modèles géométriques. Le principe de ces modèles est de considérer les relations comme des transformations géométriques dans l'espace vectoriel et d'utiliser une distance faisant intervenir les plongements des nœuds et des relations comme fonction de score.

Le premier modèle de ce type, `TransE` (Bordes et al., 2013) considère qu'un triplet est correct si la somme des plongements de l'entité-source et de la relation est proche du plongement de l'entité-cible, utilisant ainsi la distance euclidienne comme fonction score :

$$f(h, r, t) = -\|\mathbf{h} + \mathbf{r} - \mathbf{t}\|_2$$

où $\mathbf{h}$, $\mathbf{r}$ et $\mathbf{t}$ représentent respectivement les plongements de l'entité-source, de la relation et de l'entité-cible.

Plusieurs modèles se sont inspirés de `TransE`, lançant alors la série de modèles `Trans` avec `TransH` (Wang et al., 2014), `TransR` (Lin et al., 2015) et `TransD` (Ji et al., 2015) offrant diverses extensions.

D'autres modèles, comme `RotatE` (Sun et al., 2019), utilisent des $\mathbb{C}$-espaces vectoriels pour représenter les entités et définissent les relations comme des rotations entre elles :

$$f(h, r, t) = -\|\mathbf{h} \circ \mathbf{r} - \mathbf{t}\|$$

où $\circ$ désigne le produit élément par élément. Ce modèle permet ainsi de modéliser des modèles de relations plus complexes comme la symétrie, l'asymétrie, les inversions et les compositions.

Modèles *semantic-matching*. Contrairement aux modèles géométriques, les modèles *semantic-matching* utilisent une fonction de score dérivée de la similarité sémantique pour découvrir les connexions sémantiques potentielles entre les entités et les relations.

Parmi les exemples notables, nous pouvons citer `RESCAL` (Nickel et al., 2011), un modèle bilinéaire qui associe à chaque relation r une matrice $\mathbf{R}_r \in \mathbb{R}^{d \times d}$, où d est la dimension de l'espace latent :

$$f(h, r, t) = \mathbf{h}^\top \mathbf{R}_r \mathbf{t}$$

Dans ce modèle, chaque relation est considérée comme une transformation linéaire spécifique d'un espace latent, transformation expressive mais coûteuse en paramètres.

Yang et al. (2015) proposent alors `DistMult` qui réduit la charge de calcul et simplifie `RESCAL` en imposant une matrice $\mathbf{R}_r$ diagonale (taille de $\mathbf{R}_r$ à $\text{diag}(\mathbf{r}) : d^2 \to d$) :

$$f(h, r, t) = \mathbf{h}^\top \text{diag}(\mathbf{r}) \mathbf{t}$$

Chaque dimension du plongement d'une entité contribue indépendamment au score, avec un poids donné par la relation r. Cependant cette modélisation est symétrique par construction, la rendant inadaptée pour modéliser l'asymétrie.

Le modèle `ComplEx` (Trouillon et al., 2016) généralise `DistMult` à l'espace complexe permettant une capacité de modélisation plus large et la modélisation des relations asymétriques, tout en gardant un nombre de paramètres de l'ordre de d.

Modèles profonds. Les approches présentées précédemment se limitent à l'utilisation d'opérations mathématiques simples, telles que le produit scalaire ou les multiplications de matrices sur les plongements d'entités et de relations. Par conséquent, leur capacité de modélisation ne peut être augmentée qu'en modifiant les dimensions des plongements.

L'application des réseaux neuronaux profonds à ces graphes constitue une alternative pour pallier ces limites. Non triviale, elle a été sous-explorée jusqu'au développement de deux types de modèles : les modèles basés sur **les réseaux de neurones convolutifs**, *Convolutional Neural Network* (CNN) et ceux basés sur **les réseaux convolutifs graphiques**, *Graph Convolutional Network* (GCN).

Les réseaux de neurones convolutifs tels que `ConvE` (Dettmers et al., 2018) et `ConvKB` (Nguyen et al., 2018) appliquent des couches de convolution sur l'espace latent pour modéliser les interactions entité-relation et utilisent une couche dense pour calculer le score.

Parallèlement, les réseaux convolutifs graphiques (GCN), proposés par Kipf and Welling (2017), ont permis de propager les informations de différents nœuds à travers les chemins du graphe, ce qui a conduit à des représentations d'entités et de relations tenant compte du voisinage.

Plus précisément, étant donné un graphe $G = (V, E, X)$, avec V l'ensemble des nœuds, E l'ensemble des arêtes et $X \in \mathbb{R}^{|V| \times d_0}$ les *features* d_0-dimensionnelles de chaque nœud, la représentation vectorielle des nœuds obtenue après une couche de GCN est la suivante :

$$\boldsymbol{H} = f(\hat{A}XW)$$

où $\hat{A}$ est la matrice d'adjacence normalisée de G avec boucles (arêtes vers soi-même) ajoutées, $W \in \mathbb{R}^{d_0 \times d_1}$ sont les paramètres de la couche et f une fonction d'activation. La représentation $\boldsymbol{H}$ encode le voisinage immédiat de chaque nœud. Il est possible d'ajouter plusieurs couches, ainsi pour la couche k :

$$\boldsymbol{H}^{k+1} = f(\hat{A}\boldsymbol{H}^k W^k)$$

où $W^k \in \mathbb{R}^{d_k \times d_{k+1}}$ sont les paramètres de la couche k, avec $H^0 = X$.

Des modèles comme `R-GCN` (Schlichtkrull et al., 2018) et `CompGCN` (Vashishth et al., 2020) proposent d'autres modifications pour traiter les graphes multi-relationnels.

3.4.4.3 Les méthodes basées sur les grands modèles de langue

De nouvelles méthodes intégrant des modèles de langue et particulièrement les grands modèles de langue apparaissent dernièrement pour résoudre la tâche de prédiction de liens.

Bi et al. (2024) proposent le *framework* `LPNL` (*Link Prediction via Natural Language*) qui se base sur le modèle `T5` affiné et des *prompts* décrivant le graphe pour la prédiction de lien à grande échelle sur des graphes hétérogènes.

Dans la même veine, Shu et al. (2024) introduisent un autre *framework*, `KG-LLM`, qui exploite les grands modèles de langue pour résoudre des tâches sur les graphes de connaissances, dont la prédiction de liens. La structure du graphe est décrite avec une stratégie de *Chain-of-Thought* et trois grands modèles de langue sont affinés pour traiter des *prompts* inconnus en *zero-shot*. Leur approche traite notamment la prédiction de liens *multi-hops*, i.e. qui nécessite plusieurs étapes de raisonnement. Par exemple, soient les triplets `(Alice, travailler_à, EntrepriseX)`, `(EntrepriseX, situé_dans, Paris)`, `(Paris, capitale, France)` dans un graphe de connaissance, le modèle peut compléter `(Alice, vivre_en, France)` avec un raisonnement en trois *hops* (étapes).

Des approches hybrides combinant les grands modèles de langue et les réseaux de neurones graphiques (*Graph Neural Networks*) (`GNN`) pour la prédiction de liens apparaissent également. Xiao et al. (2024) proposent `FuseLinker`, système qui intègre des plongements de textes pré-entraînés issus de grands modèles de langues et apprend la structure du graphe à partir de `GNN`. Cette approche hybride permet d'améliorer les performances de prédiction des liens dans les graphes de connaissances biomédicaux en injectant des caractéristiques textuelles sémantiques spécifiques au domaine.

3.5 Conclusion

Dans ce chapitre, nous avons discuté de la modélisation de l'information sous forme de graphe, conduisant à la construction de graphes de connaissances ou de réseaux lexico-sémantiques selon le type d'informations décrites (informations du domaine général ou linguistiques). Ces graphes ont fait l'objet de nombreux travaux en TAL en raison de leur capacité à décrire des informations interconnectées de manière structurée. Ils présentent toutefois la limite d'être intrinsèquement incomplets.

Si la question de l'exhaustivité et de l'incomplétude se pose pour l'ensemble des ressources, elle a été particulièrement étudiée dans le cadre des graphes de connaissances et des réseaux lexico-sémantiques avec le développement de différentes méthodes et de tâches pour la complétion.

Nous avons présenté ces méthodes, en distinguant celles dite internes, qui exploitent les informations existantes du graphe, et celles dites externes, qui s'appuient sur des sources d'information extérieures pour enrichir la ressource.

Plus spécifiquement, nous nous sommes concentrés sur la complétion à travers la tâche de prédiction de liens, qui constitue une tâche majeure dans le domaine de la complétion des graphes de connaissances (*Knowledge Graph Completion*). Nous avons introduit la tâche et les jeux de données principaux construits à partir de graphes de connaissances et les principaux systèmes, avec une attention particulière aux systèmes basés sur les plongements de graphe.

Si les méthodes de complétion externe, en particulier celles qui combinent l'extraction *web*, la supervision à distance ou l'enrichissement manuel, montrent une augmentation concrète du nombre de triplets dans le graphe, les travaux sur la prédiction de liens se concentrent davantage sur la construction des systèmes et la manière de représenter les informations structurelles du graphe pour inférer des informations, qui sont artificiellement retirées pour le *design* de la tâche. Par conséquent, la prédiction de liens ainsi conduite n'entraîne en réalité pas d'enrichissement concret de la ressource.

Dans le chapitre 4, nous nous intéresserons à cette problématique en évaluant des prédictions d'un modèle de prédiction de liens pour extraire de nouveaux triplets potentiels.

Chapitre 4

Génération de prédictions robustes pour l'enrichissement de réseaux lexico-sémantiques

Sommaire

4.1 Enjeux et contributions 71
4.2 Réseaux lexico-sémantiques du français 73
4.2.1 De `RezoJDM` à `RezoJDM16k` 73
4.2.2 Le `RL-fr`, Réseau Lexical du Français 74
4.2.3 Deux réseaux lexico-sémantiques différents mais complémentaires . 75
4.3 Prédiction de liens 76
4.3.1 Pré-traitement des jeux de données 76
4.3.2 Modèles et métriques 77
4.3.3 Comparaison des résultats 78
4.4 Robustesse du modèle à travers l'incertitude prédictive 80
4.4.1 Conception et entraînement d'un réseau de neurones profond 80
4.4.2 Modèle formel de l'incertitude prédictive 81
4.4.3 Approximation de l'incertitude avec le Monte-Carlo Dropout 82
4.5 Enrichir un réseau lexico-sémantique parcimonieux 84
4.5.1 Application du Monte-Carlo Dropout 84
4.5.2 Sélection de triplets candidats 84
4.5.3 Annotations manuelles 85
4.5.4 Extraction de nouveaux triplets 87
4.6 Conclusion 88

4.1 Enjeux et contributions

Dans le chapitre 3, nous avons présenté un état de l'art sur la complétion des graphes de connaissances et les réseaux lexico-sémantiques *via* la tâche de prédiction de liens transductive. Étant donné un graphe représenté sous un ensemble de triplets `(h, r, t)` pour `(entité-source, relation, entité-cible)`, la tâche consiste à prédire correctement l'entité manquante d'un triplet incomplet, par exemple `(entité-source, relation, ?)` ou `(?, relation, entité-cible)`.

Bien que la nécessité de compléter les graphes de connaissances qui sont incomplets soit un argument avancé dans la majorité des travaux en prédiction de liens, nous remarquons que ces travaux entraînent rarement un réel enrichissement de la ressource. En effet, en pratique, les systèmes de prédiction de liens sont évalués sur des données existantes du graphe qui ont été retirées artificiellement pour constituer l'échantillon d'évaluation. Ce processus est classique pour l'évaluation des performances des modèles mais ne donne pas lieu à de nouveaux triplets permettant d'enrichir le graphe.

Dans ce chapitre, nous décidons d'adopter une approche davantage orientée vers la ressource, en exploitant les prédictions générées par des modèles de prédiction de liens, dans le but d'enrichir un réseau lexico-sémantique parcimonieux, le `RL-fr`. Notre objectif est d'aller au-delà des performances des modèles sur la tâche et de proposer une méthode permettant de générer des prédictions robustes, susceptibles de constituer de nouvelles informations pour le graphe. La configuration transductive de la tâche implique que des nouvelles entités et/ou relations ne peuvent pas apparaître, toutefois, des connexions manquantes peuvent être inférées par les modèles.

Ce chapitre est ainsi consacré à la description de nos expériences sur deux réseaux lexico-sémantiques du français opposés : `RezoJDM`, un réseau très dense construit semi-automatiquement par myriadisation et le `RL-fr`, un réseau relativement parcimonieux car conçu manuellement par des lexicographes. Utiliser ces deux ressources nous permet de comparer les performances des modèles sur des graphes de densité différente.

Nous proposons les contributions suivantes dans cette partie de la thèse :

1. Nous évaluons les performances de sept modèles de prédiction de liens sur deux réseaux lexico-sémantiques du français, dont un modèle présentant des résultats état-de-l'art sur un jeu de données existant.

2. Nous exploitons la technique de Monte-Carlo Dropout (Gal and Ghahramani, 2016) pour modéliser l'incertitude du modèle et générer des prédictions robustes par le modèle le plus performant, `CompGCN-ConvE`.

3. Nous analysons la mesure de robustesse à partir d'annotations manuelles, montrant une corrélation entre un score de confiance élevé et des triplets pertinents pour le `RL-fr`.

Le chapitre s'articule autour de plusieurs sections : la section 4.2 présente les spécificités des deux réseaux lexico-sémantiques du français `RezoJDM` et `RL-fr`, ainsi que leurs différences. Nous présentons dans la section 4.3, le pré-traitement des données, les modèles et les métriques utilisés ainsi que les résultats des expériences.

Nous détaillons ensuite dans la section 4.4, une description théorique de l'incertitude des modèles et la technique de Monte-Carlo Dropout qui nous permet de générer des prédictions robustes avec un score de confiance. Dans la section 4.5, nous appliquons le Monte-Carlo Dropout et nous évaluons manuellement si cette technique est en mesure de prédire de nouveaux triplets potentiels pour le `RL-fr`. Pour finir, nous discutons dans la section 4.6 des limites et des conclusions que nous tirons de ces travaux.

Les expériences relatives à ce chapitre ont fait l'objet de deux publications (Choi et al., 2024b,a).

4.2 Réseaux lexico-sémantiques du français

Dans cette section, nous présentons les deux réseaux lexico-sémantiques du français que nous utilisons dans nos expériences : `RL-fr` (Lux-Pogodalla and Polguère, 2011) et `RezoJDM` (Lafourcade and Joubert, 2008; Lafourcade and Le Brun, 2020).

4.2.1 De `RezoJDM` à `RezoJDM16k`

`RezoJDM` (Lafourcade and Joubert, 2008; Lafourcade and Le Brun, 2020) est un réseau lexico-sémantique du français développé *via* des jeux ayant un but, des approches contributives et des mécanismes d'inférence.

La plateforme `JeuxDeMots` [39] propose en effet des jeux permettant de développer le réseau en ajoutant de nouvelles entrées et de vérifier des informations du réseau. Le jeu principal invite les joueurs à saisir des termes dans un délai imparti selon un terme et un type de relation donnés (cf. figure 4.1). Lancée en 2006, cette ressource a fait l'objet de mises à jour régulières. L'utilisation des jeux ayant un but a ainsi permis de créer un graphe dirigé très dense, puisqu'il comprend actuellement plus d'un milliard de relations et 15 millions de nœuds [40].

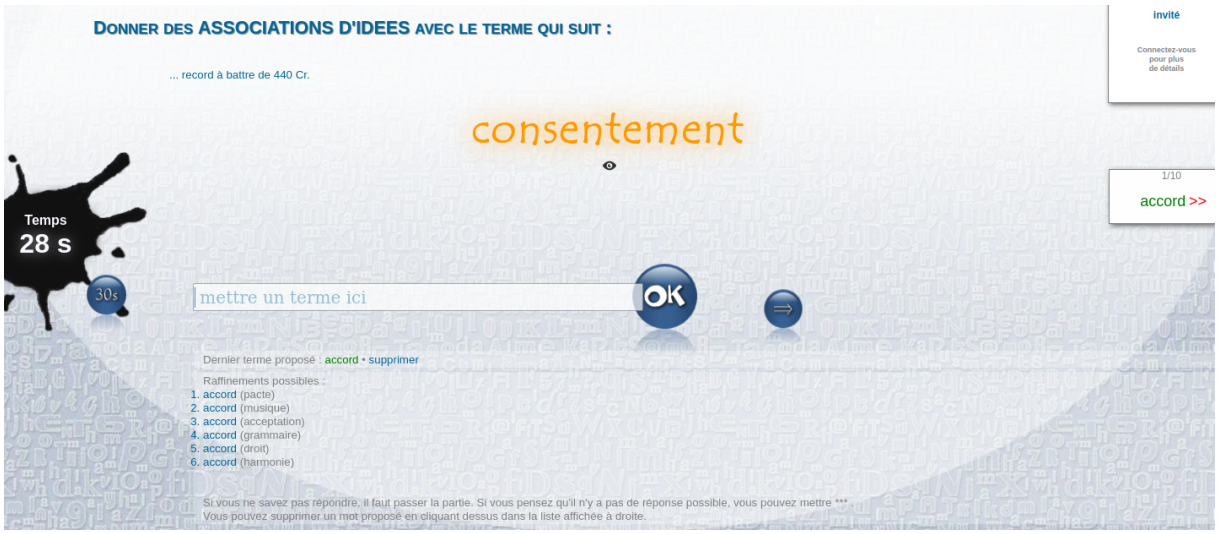


FIGURE 4.1 – Exemple d'une partie de `JeuxDeMots` où le participant doit entrer des termes associés au terme `consentement`. Le dernier terme entré est `accord` et le jeu présente ses différents sens (raffinements). Le participant peut raffiner sa réponse en choisissant un des sens proposés.

Dans `RezoJDM`, les nœuds représentent principalement des termes qui correspondent à des lemmes (type `n_term`) mais ils peuvent également décrire des informations telles que des étiquettes de partie du discours (type `n_pos`) ou des formes fléchies (type `n_form`). Les relations sont divisées en trois catégories : lexicales (synonymie, antonymie...), ontologiques (hyperonymie, méronymie...) et prédicatives (agent, conséquences...).

Les nœuds et les relations ont la particularité de présenter des poids en fonction de la dynamique du jeu. Pour les relations, un poids positif code une relation vraie et un poids négatif code une relation fausse, attribut relativement rare dans les graphes de connaissances. La valeur

39. `https://www.jeuxdemots.org`
40. En juillet 2025.

du poids est dépendante des différents jeux proposés dans JeuxDeMots.

Dans `RezoJDM`, la polysémie d'un terme est exprimée en distinguant un nœud générique de ses nœuds de raffinement. Cependant, comme le jeu demande au participant d'entrer autant de termes que possible dans un temps limité, les joueurs ont tendance à ne pas affiner leurs réponses, ce qui conduit à une faible densité des nœuds de raffinement. Par exemple, le nœud générique `accord` a un degré de 10 549, tandis que les degrés des nœuds de raffinement `accord>pacte` et `accord>acceptation` sont respectivement de 194 et 123.

En raison de son volume très important [41], `RezoJDM` se prête difficilement à une manipulation à l'échelle de l'ensemble de la ressource.

`RezoJDM16k` est un sous-graphe de `RezoJDM` créé par Mirzapour et al. (2022) pour la tâche de prédiction de liens. Il est obtenu en appliquant divers filtres aux nœuds et relations de `RezoJDM`. Seuls les nœuds de type `n_term` et de poids supérieur à 50 ont été retenus. Le même filtre de poids a été appliqué aux relations et certains types relatifs aux métadonnées ont également été supprimés.
En outre, les types de relations apparaissant moins de 100 fois et les nœuds ayant un degré inférieur à 45 ont été exclus, afin d'améliorer l'efficacité des modèles. Le graphe final est composé de 15 746 nœuds et de 832 093 relations.

4.2.2 Le `RL-fr`, Réseau Lexical du Français

Le Réseau Lexical du Français `RL-fr` est un réseau lexico-sémantique du français créé par des lexicographes, où les nœuds correspondent à des unités lexicales et les arêtes à des relations lexico-sémantiques ou combinatoires (Lux-Pogodalla and Polguère, 2011). Dans cette section, nous présentons des informations concernant la version 2.1 de `RL-fr`, utilisée pour nos expériences [42].

Les unités lexicales dans `RL-fr` sont les entités fondamentales pour la description lexicographique et peuvent être soit un lexème, soit un idiome. Les lexèmes sont des unités lexicales monosémiques et correspondent aux sens des mots. Par conséquent, un lexème polysémique, appelé vocable, est représenté comme une collection d'unités lexicales interconnectées par une relation de copolysémie. La version 2.1 du `RL-fr` contient 29 220 unités lexicales et 18 625 vocables. Voici les lexèmes du vocable `jambe` illustrés par des exemples lexicographiques tirés du `BEL-RL-fr` (Lux-Pogodalla, 2014), à titre d'exemple :

— `Jambe I.1` : Marc attend patiemment, les **jambes** croisées.
— `Jambe I.2a` : Le cheval s'est blessé à la **jambe**.
— `Jambe I.2b` : Il y a de la **jambe** de porc au menu.
— `Jambe II` : La **jambe** droite du pantalon est déchirée.
— `Jambe III` : Une des **jambes** de suspension doit être changée.

Contrairement aux dictionnaires classiques, où les sens d'un mot sont généralement seulement listés, le `RL-fr` représente les types de relation entre les sens avec la notion de copolysémie,

41. Un *dump* extrait en 2022 contenant 17 148 882 nœuds et 406 355 459 relations faisait 18 Go.
42. `https://www.ortolang.fr/market/lexicons/lexical-system-fr/v2.1`

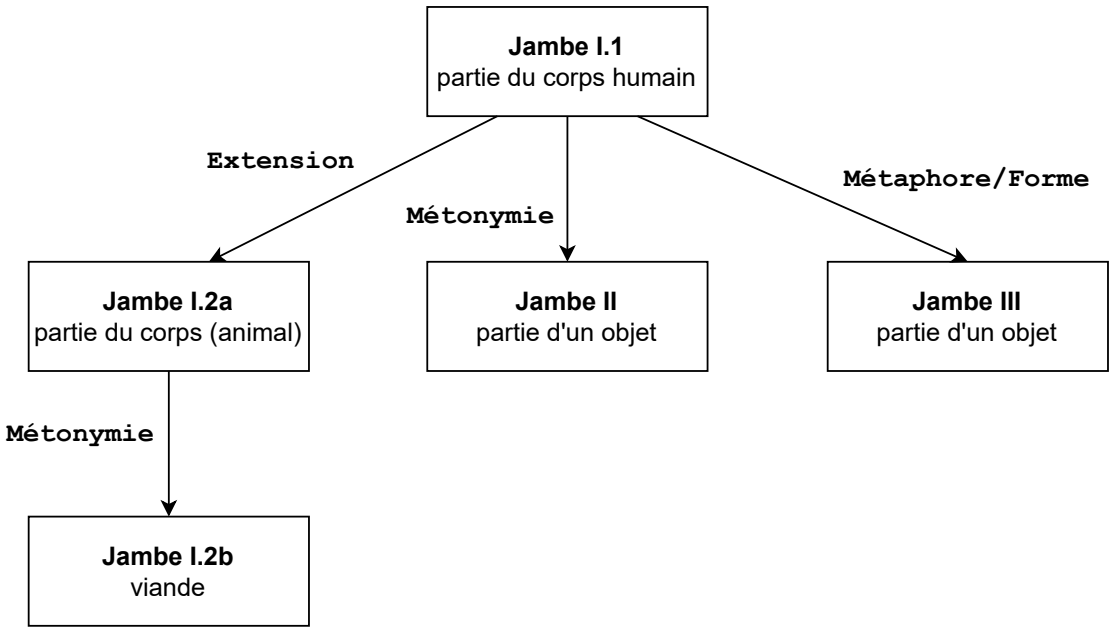


FIGURE 4.2 – Représentation de la copolysémie dans le RL-fr.

décrite comme la relation entre les différents sens d'un mot, opposée à la polysémie qui est la propriété des mots d'exprimer plusieurs sens (Polguère, 2018). Il existe donc plusieurs relations de copolysémie, telle que la métonymie ou la métaphore.

La ressource est basée sur la théorie Sens-Texte qui permet d'encoder les relations paradigmatiques et syntagmatiques avec des fonctions lexicales (Mel'čuk, 1996) (cf. Section 1.3.3). Elles peuvent être utilisées pour représenter des relations sémantiques simples ou complexes.

Dans la version 2.1 du RL-fr, il existe 686 fonctions lexicales différentes. Pour nos expériences, nous avons utilisé les familles des relations plutôt que les relations elles-mêmes, ce qui correspond déjà à 95 types de fonctions lexicales et 11 types de relations de copolysémie. Au total, il existe 62 641 relations encodées par des fonctions lexicales et 9 413 relations de copolysémie.

4.2.3 Deux réseaux lexico-sémantiques différents mais complémentaires

Bien qu'ils partagent le même format de réseau, RezoJDM16k et le RL-fr présentent des différences significatives en termes de création, de portée et de représentation de la polysémie. Le tableau 4.1 illustre le contraste entre ces deux graphes : RezoJDM16k est très dense, avec un degré moyen de nœuds de 105,7 et un nombre d'arêtes plus de 10 fois supérieur à celui de RL-fr. Par ailleurs, RL-fr présente davantage de nœuds, en raison de la représentation des sens des mots dans des nœuds distincts. Il est toutefois plus parcimonieux que le RezoJDM16k en raison de sa construction complètement manuelle.

La figure 4.3 montre un aperçu du réseau RL-fr autour du vocable accord, représenté par deux nœuds accord 1 et accord 2. La figure a été réalisée avec spiderlex [43].

43. https://spiderlex.atilf.fr/fr/q/*accord***

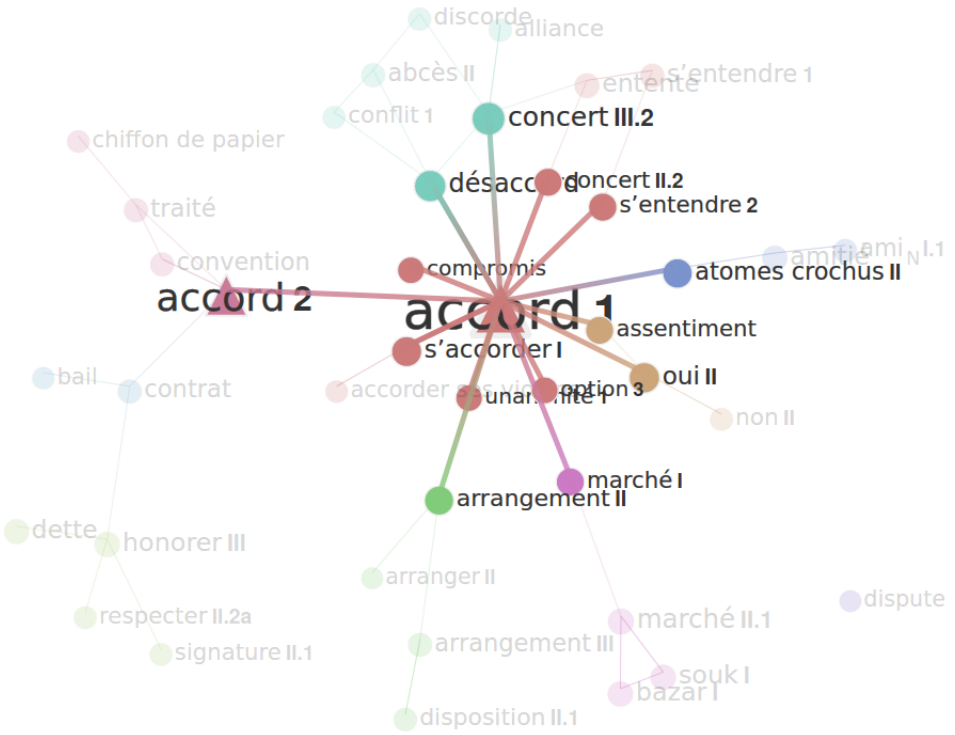


FIGURE 4.3 – Vocable `accord` représenté dans le `RL-fr`.

	`RezoJDM16k`	`RL-fr`
# nœuds	15 746	29 220
# arêtes	832 093	72 054
# types d'arête	150	106
Degré moyen d'un nœud	105,7	5,2
Degré minimum	2	1
Degré maximum	10 403	187

TABLE 4.1 – Statistiques sur `RezoJDM16k` et `RL-fr`.

4.3 Prédiction de liens

4.3.1 Pré-traitement des jeux de données

Toutes les expériences sont réalisées dans une configuration transductive, ce qui implique que tous les nœuds à prédire sont connus de l'entraînement. Pour respecter cette condition, Mirzapour et al. (2022) utilisent un algorithme vérifiant si le nœud source et le nœud cible de chaque triplet sont impliqués dans au moins un autre triplet. Lorsque cette condition est vérifiée, le triplet est placé dans l'échantillon de validation. Une fois que la taille souhaitée de l'échantillon de validation est atteinte, le même processus est répété sur les triplets restants afin d'obtenir les échantillons de test et d'entraînement.

Bien que cela garantisse l'inclusion de tous les nœuds du graphe dans l'échantillon d'entraînement, cette méthode n'est pas adaptée à des graphes parcimonieux tels que le `RL-fr` car le

nombre de triplets validant la condition de l'algorithme de sélection n'est pas suffisant. La taille souhaitée n'est donc jamais atteinte.

Par conséquent, nous décidons de diviser aléatoirement les triplets entre les échantillons d'entraînement, de validation et de test dans un ratio 80 % - 10 % - 10 % et de vérifier si tous les nœuds des échantillons de validation et de test sont présents dans l'entraînement. Dans le cas contraire, nous supprimons le triplet de l'échantillon.

Pour un graphe très dense comme `RezoJDM16k`, aucun triplet n'a été perdu même après dix divisions aléatoires. En revanche, un graphe parcimonieux tel que le `RL-fr` subit des pertes d'arêtes et de nœuds (- 2 152 nœuds, - 1 037 arêtes). Nous obtenons ainsi un graphe final de 27 068 nœuds et 71 017 arêtes, que nous nommons `RLF27k`. Le tableau 4.2 présente une comparaison des jeux de données.

	`RezoJDM16k`	`RLF27k`
# nœuds	15 746	27 068
# arêtes	832 093	71 017
# triplets entraînement	665 674	57 643
# triplets validation	83 209	6 674
# triplets test	83 210	6 700

TABLE 4.2 – Statistiques sur les échantillons d'entraînement, validation et test de `RezoJDM16k` et `RLF27k` après une division en 80 %, 10 %, 10 %.

4.3.2 Modèles et métriques

Modèles. Nous choisissons d'utiliser les six modèles présents dans Mirzapour et al. (2022) afin de reproduire les expériences sur `RezoJDM16k` et d'observer les résultats sur `RLF27k`. Les modèles ont été utilisés *via* la librairie OpenKE (Han et al., 2018).

Afin d'explorer les performances d'un modèle profond sur des données en français, nous ajoutons le modèle `CompGCN-ConvE`, composé du modèle `ConvE` et de l'encodeur `CompGCN` (Vashishth et al., 2020). Le tableau 4.3 présente les caractéristiques principales des modèles utilisés (pour plus de détails, cf. Section 3.4.4).

Métriques Nous utilisons les trois métriques traditionnelles détaillées en section 3.4.3.

— **Mean Rank (MR) :** étant donné un ensemble de triplets classés par leur score de prédiction, le MR calcule le rang moyen des triplets corrects. Un MR faible signifie une meilleure performance.

— **Mean Reciprocal Rank (MRR) :** MRR est la moyenne des inverses des rangs des triplets corrects. Un MRR élevé indique une meilleure performance.

— **Hits@k :** cette mesure calcule la proportion de triplets corrects apparaissant dans le top k de la liste classée des triplets prédits. Une valeur Hits@k élevée indique une précision de

Modèle	Fonction score $f(h, r, t)$	Caractéristiques
`TransE` (Bordes et al., 2013)	$-\|\mathbf{h} + \mathbf{r} - \mathbf{t}\|_{L_2}$	Premier modèle géométrique type `Trans`
`TransH` (Wang et al., 2014)	$\| (\mathbf{h} - \mathbf{w}_r^\top h\, \mathbf{w}_r) + \mathbf{d}_r - (\mathbf{t} - \mathbf{w}_r^\top t\, \mathbf{w}_r) \|_{L_2}$	Projette les plongements dans l'hyperplan $\mathbf{w}_r$ défini par la relation $\mathbf{r}$
`TransD` (Ji et al., 2015)	$-\|\mathbf{M}_{rh}\mathbf{h} + \mathbf{r} - \mathbf{M}_{rt}\mathbf{t}\|_{L_2}$, $\mathbf{M}_{rh} = \mathbf{r}_p\mathbf{h}_p + \mathbf{I}^{m\times n}$, $\mathbf{M}_{rt} = \mathbf{r}_p\mathbf{t}_p + \mathbf{I}^{m\times n}$	Projette chaque plongement différemment, selon à la relation $\mathbf{r}$
`RotatE` (Sun et al., 2019)	$-\|\mathbf{h} \circ \mathbf{r} - \mathbf{t}\|_{L_2}$	Modélise la symétrie, l'asymétrie, l'inverse et la composition
`DistMult` (Yang et al., 2015)	$\mathbf{h}^\top \mathrm{diag}(\mathbf{r})\mathbf{t} = \sum_i h_i r_i t_i$	Peu de paramètres, mais impossible de modéliser des relations asymétriques
`ComplEx` (Trouillon et al., 2016)	$\Re\left(\langle \mathbf{h}, \mathbf{r}, \overline{\mathbf{t}} \rangle\right) = \Re\left(\sum_i h_i\, r_i\, \overline{t_i}\right)$	Étend `DistMult` aux complexes, modélise l'asymétrie en restant léger en paramètres
`CompGCN-ConvE` (Vashishth et al., 2020)	$\sigma(\mathrm{vec}(\mathrm{Conv2D}(\mathrm{concat}(\mathbf{h}, \mathbf{r})))\, \mathbf{W}) \cdot \mathbf{t}$	Combine un modèle de convolution de graphe avec un décodeur convolutionnel

TABLE 4.3 – Description des modèles de prédiction de liens utilisés. $\Re$ est l'opérateur « partie réelle ».

prédiction supérieure pour les triplets les plus importants.

4.3.3 Comparaison des résultats

Les résultats pour `RezoJDM16k` sont présentés dans le tableau 4.4. Nous constatons que `CompGCN-ConvE` surpasse tous les autres modèles avec pratiquement toutes les métriques. Il atteint notamment un Hits@1 de 0,357, qui est plus de deux fois supérieur à celui de `RotatE`. Les scores Hits@3, Hits@10 et MRR sont également les plus élevés et le MR reste proche du meilleur score obtenu par `TransD`.

Le tableau 4.5 présente les résultats pour `RLF27k`. Une fois de plus, `CompGCN-ConvE` se distingue avec un MRR de 0,515 et les meilleurs scores Hits@k. Cependant, nous pouvons noter que le MR élevé indique une disparité dans le classement des triplets corrects : 60 % d'entre eux se classent parmi les dix premiers, tandis que les 40 % restants sont nettement moins bien classés.

Modèle	**MRR ↑**	**MR ↓**	**Hits@10 ↑**	**Hits@3 ↑**	**Hits@1 ↑**
`TransE`	0,180	200,78	0,437	0,242	0,040
`TransH`	0,217	173,28	0,503	0,293	0,064
`TransD`	0,216	**168,18**	0,500	0,290	0,065
`DistMult`	0,219	194,16	0,446	0,252	0,109
`ComplEx`	0,256	190,79	0,539	0,309	0,119
`RotatE`	0,312	177,04	0,587	0,409	0,155
`CompGCN-ConvE`	**0,461**	171,26	**0,659**	**0,514**	**0,357**

TABLE 4.4 – Résultats des modèles de prédiction de liens sur `RezoJDM16k`.

Modèle	MRR ↑	MR ↓	Hits@10 ↑	Hits@3 ↑	Hits@1 ↑
`TransE`	0,278	**2 594,24**	0,624	0,497	0,033
`TransH`	0,250	2 957,59	0,581	0,465	0,011
`TransD`	0,255	2 752,03	0,587	0,472	0,016
`DistMult`	0,373	2 748,25	0,613	0,502	0,216
`ComplEx`	0,413	3 447,98	0,593	0,524	0,284
`RotatE`	0,399	3 650,92	0,490	0,454	0,336
`CompGCN-ConvE`	**0,515**	2 808,68	**0,627**	**0,559**	**0,450**

TABLE 4.5 – Résultats des modèles de prédiction de liens sur `RLF27k`.

Lors de l'inférence, les modèles classent les prédictions pour une entité et une relation données selon leur score. La figure 4.4 présente les 20 meilleures prédictions pour la tête `bonnet I` et la relation `synonymie` de `RLF27k` avec le meilleur modèle, `CompGCN-ConvE`. Nous précisons également si les triplets existent déjà dans `RLF27k`, et le cas échéant, l'échantillon correspondant.

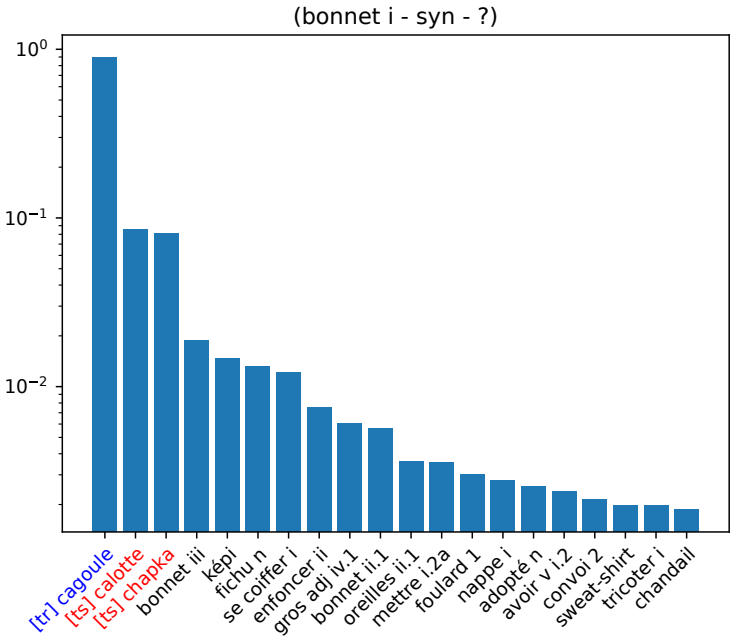


FIGURE 4.4 – Prédictions pour la tête `bonnet I` et la relation `synonymie`. Les triplets existant déjà dans le `RLF27k` sont précisés avec les couleurs : en bleu, ceux dans l'échantillon d'entraînement ([tr] pour *training*) et en rouge, ceux dans l'échantillon de test ([ts] pour test). L'axe des ordonnées est mis à l'échelle logarithme pour une meilleure lisibilité.

En premier lieu, le modèle prédit logiquement un triplet présent dans l'entraînement avec un score élevé à 0,893, puis prédit deux triplets du test avec des scores d'environ 0,08.

Nous supposons que les triplets restants, qui n'existent pas dans le graphe original, pourraient être de potentiels nouveaux triplets. Néanmoins, la fonction score seule ne permet pas d'évaluer significativement la pertinence de ces triplets, les résultats obtenus étant particulièrement faibles, entre 0,0189 et 0,0019.

Par ailleurs, un modèle entraîné génère un ensemble fixé de prédictions mais peut produire des prédictions complètement différentes s'il est entraîné avec d'autres paramètres. De fait, nous ne capturons pas **l'incertitude du modèle** et nous ne savons pas si les prédictions basées sur une « version » du modèle sont robustes.

4.4 Robustesse du modèle à travers l'incertitude prédictive

Dans cette section, nous nous proposons d'expliciter la modélisation théorique de l'incertitude, telle que définie par Gal and Ghahramani (2016) et Gawlikowski et al. (2023).

Dans un premier temps, nous modélisons la conception et l'entraînement d'un réseau de neurones profond et son incertitude intrinsèque en sections 4.4.1 et 4.4.2.

Dans un second temps, nous expliquons en section 4.4.3 que cette incertitude est généralement ignorée dans les réseaux de neurones profonds standards. Toutefois, elle peut être approximée en pratique avec le Monte-Carlo Dropout, en s'inspirant des réseaux de neurones Bayésiens.

4.4.1 Conception et entraînement d'un réseau de neurones profond

Un réseau de neurones profond peut être formellement défini comme une **fonction non-linéaire**

$$f_{\boldsymbol{\theta}} : \mathbb{X} \to \mathbb{Y}$$

paramétrée par un ensemble de paramètres $\boldsymbol{\theta}$ (par exemple, poids et biais). Ici, $\mathbb{X}$ est l'espace des entrées et $\mathbb{Y}$ est l'espace des sorties.

En apprentissage supervisé, on dispose d'un ensemble fini de données d'entraînement

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N} \subset \mathbb{X} \times \mathbb{Y}$$

contenant N échantillons et leurs cibles correspondantes.

Étant donné une nouvelle entrée $x^* \in \mathbb{X}$, on cherche à prédire la sortie $y^* \in \mathbb{Y}$ correspondante. Soit $x^* \in \mathbb{X}$, un réseau entraîné avec $\mathcal{D}$ peut être utilisé pour produire une prédiction :

$$f_{\boldsymbol{\theta}}(x^*) = y^*$$

Cette prédiction peut cependant contenir de l'incertitude, présente lors de la capture des données, de la construction du modèle et de son entraînement, ou de l'inférence. Ces incertitudes proviennent :

- de bruit dans les données (bruit lors de la captation, erreur humaine d'annotation) ;
- d'erreurs inhérentes aux systèmes de mesures (erreur de calibration d'une caméra par exemple) ;
- d'erreurs dans le choix d'architecture du modèle (modèle peu expressif (peu de paramètres), architecture inadaptée à la tâche) ;
- d'erreurs lors de l'entraînement du modèle (mauvais hyperparamètres, mauvaise initialisation des paramètres) ;
- d'erreurs causées par des données inconnues (par exemple lors de l'inférence avec des données hors du vocabulaire du modèle ou de nouvelles classes).

La conception d'un réseau de neurones profond inclut le choix de l'architecture du modèle ainsi que son entraînement, qui est un procédé stochastique.

Nous résumons l'architecture du modèle par une configuration s (par exemple, nombre de paramètres, couches, fonctions d'activation). s représente l'architecture du modèle, fixée avant l'entraînement. $\boldsymbol{\theta}$ représente la valeur numérique des poids du modèle, qui sont appris lors de l'entraînement.

Étant donné $(s, \mathcal{D})$, l'entraînement d'un modèle est un processus stochastique en raison de la présence potentielle d'aléas dans l'ordre des données lors du *mini-batching*, l'initialisation aléatoire des paramètres ou encore la régularisation aléatoire du *dropout* ou de l'augmentation.

De plus, le fait que la surface de la fonction de perte du modèle soit non convexe conduit souvent à des optima locaux $\boldsymbol{\theta}^*$ qui peuvent être différents selon l'entraînement, même avec des s et $\mathcal{D}$ identiques. Des hyperparamètres tels que la taille des *batchs*, le *learning rate* et le nombre d'*epoch* influencent également le modèle final, entraînant des variations dans les prédictions.

Ces sources d'incertitudes impliquent qu'un ensemble de modèles $\{f^i_{\boldsymbol{\theta}}\}$ entraînés avec les mêmes s et $\mathcal{D}$ peuvent produire, pour un x^* donné, des prédictions $\{f^i_{\boldsymbol{\theta}}(x^*) = y^*_i\}$ potentiellement différentes.

4.4.2 Modèle formel de l'incertitude prédictive

Soit $x^* \in \mathbb{X}$ une nouvelle entrée et $y^* \in \mathbb{Y}$ son étiquette. On s'intéresse à la distribution prédictive :

$$p(y^* \mid x^*)$$

Il s'agit de la distribution de probabilités sur les sorties possibles y^* pour une nouvelle entrée x^*, étant données toutes les informations dont nous disposons (généralement les données d'entraînement $\mathcal{D}$ et notre modèle).

En présence d'une variable latente ω (par exemple, les paramètres du modèle $\boldsymbol{\theta}$) appartenant à l'ensemble des valeurs possibles Ω, la loi des probabilités totales donne :

$$p(y^* \mid x^*) = \int_{\Omega} p(y^* \mid x^*, \omega)\, p(\omega \mid x^*)\, d\omega$$

Avec :

$p(y^* \mid x^*, \omega)$ **:** la vraisemblance de la sortie y^* sachant l'entrée x^*, en supposant que les paramètres ω soient connus. Dans le contexte d'un réseau de neurones, cela correspond simplement au *forward* du réseau avec les paramètres ω. Celle-ci est calculable directement lorsque ω est connu.

$p(\omega \mid x^*)$ **:** la distribution *a posteriori* sur les paramètres ω sachant uniquement la nouvelle entrée x^*. Ce qui signifie : « Étant donné que l'entrée est x^*, quelle est la probabilité de chaque configuration possible des paramètres ω ? ».

Cette probabilité est incalculable car le véritable processus générateur de x^* est inconnu, ce qui empêche un calcul exact.

$\int \cdots d\omega$ **:** l'intégration sur toutes les configurations possibles des paramètres, pondérées par leur probabilité selon $p(\omega \mid x^*)$. Elle est incalculable, dans les espaces de paramètres de grande dimension, ce qui est typique des réseaux de neurones. Une intégration directe est prohibitive en coût de calcul.

En pratique, comme $p(\omega \mid x^*)$ est incalculable, on le remplace par une approximation issue de $\mathcal{D}$ (la distribution prédictive bayésienne) et on suppose que :

$$p(y^* \mid x^*) \approx \int_\Omega p(y^* \mid x^*, \omega)\, p(\omega \mid \mathcal{D})\, d\omega$$

Qui est souvent résumé dans la littérature (Gawlikowski et al., 2023) par :

$$p(y^* \mid x^*) \approx p(y^* \mid \mathcal{D}, x^*)$$

Ce qui signifie la « distribution prédictive de y^* étant donné x^* et les informations de $\mathcal{D}$ ».

Nous devons donc estimer $p(y^* \mid \mathcal{D}, x^*)$, qui est également inconnu. Il se trouve que les réseaux de neurones sont de puissants approximateurs universels capables d'estimer cette distribution prédictive à partir des données $\mathcal{D}$.

La prédiction d'un réseau de neurones est sujette à la fois à des erreurs dépendantes du modèle et à des erreurs dépendantes des données d'entrée. Par conséquent, l'incertitude prédictive associée à y^* est en général décomposée en deux types d'incertitudes :

— l'**incertitude aléatoire** (ou statistique), qui est intrinsèque aux données

— l'**incertitude épistémique** (ou systémique), qui est intrinsèque au modèle

L'approximation de $p(y^* \mid \mathcal{D}, x^*)$ avec des réseaux de neurones peut se faire de différentes manières. Nous présentons ici la méthode par Monte-Carlo (MC) Dropout, que nous utilisons dans nos expériences.

4.4.3 Approximation de l'incertitude avec le Monte-Carlo Dropout

4.4.3.1 Omission de l'incertitude par les réseaux de neurones standards

Dans un réseau de neurones standard (non bayésien), l'entraînement sur $\mathcal{D}$ produit un unique jeu de paramètres $\hat{\omega}$ (par exemple, obtenu *via* une descente de gradient stochastique). La distribution prédictive est alors approximée par :

$$p(y^* \mid \mathcal{D}, x^*) \approx p(y^* \mid x^*, \hat{\omega})$$

Cette approche suppose implicitement que les paramètres ω sont parfaitement connus après l'entraînement, ce qui revient à **ignorer l'incertitude épistémique**.

Si le réseau possède une couche de sortie probabiliste, comme par exemple, un *softmax* pour la classification ou une vraisemblance gaussienne pour la régression, la sortie $p(y^* \mid x^*, \hat{\omega})$ est généralement interprétée comme représentant l'incertitude aléatoire, c'est-à-dire l'incertitude due au bruit intrinsèque des données.

4.4.3.2 Mesure de l'incertitude avec les réseaux de neurones bayésiens

Afin de mesurer les deux types d'incertitude, les **réseaux de neurones bayésiens** considèrent les paramètres ω comme des variables aléatoires dont la distribution *a posteriori* est donnée par $p(\omega \mid \mathcal{D})$. La distribution prédictive s'écrit alors :

$$p(y^* \mid \mathcal{D}, x^*) = \int_\Omega p(y^* \mid x^*, \omega)\, p(\omega \mid \mathcal{D})\, d\omega$$

Cependant, cette intégrale est **incalculable** pour deux raisons :

— la distribution *a posteriori* $p(\omega \mid \mathcal{D})$ ne peut pas être calculée pour des réseaux profonds, car elle nécessite d'évaluer la vraisemblance pour toutes les configurations possibles des paramètres ω dans un espace de très grande dimension (de l'ordre des millions de paramètres) ;

— l'intégration elle-même sur cet espace des paramètres est impossible à effectuer en pratique.

4.4.3.3 Technique du Monte-Carlo Dropout

Le Monte-Carlo Dropout (MC Dropout) (Gal and Ghahramani, 2016) est une méthode pratique pour approximer l'inférence bayésienne dans les réseaux de neurones profonds, sans modifier l'architecture ou entraîner à nouveau le modèle. L'idée est d'interpréter le *dropout* appliqué aux couches du réseau comme une approximation variationnelle d'une distribution *a posteriori* sur les paramètres ω.

Pour rappel, la distribution prédictive bayésienne est :

$$p(y^* \mid \mathcal{D}, x^*) = \int_\Omega p(y^* \mid x^*, \omega)\, p(\omega \mid \mathcal{D})\, d\omega$$

Comme dans le cas des réseaux de neurones bayésiens, la distribution *a posteriori* $p(\omega \mid \mathcal{D})$ est incalculable, tout comme l'intégrale. Le MC Dropout contourne cette difficulté en approximant $p(\omega \mid \mathcal{D})$ par une distribution implicite $q_\phi(\omega)$ induite par les masques de *dropout* appliqués lors de l'inférence.

Concrètement, pendant l'évaluation, le *dropout* est activé, revenant ainsi à échantillonner un nouveau masque et donc un nouveau jeu de paramètres ω_t à chaque passe (*forward pass*). On obtient ainsi T prédictions stochastiques :

$$\hat{y}_t = f_{\omega_t}(x^*), \quad \omega_t \sim q_\phi(\omega), \quad t = 1, \ldots, T.$$

La distribution prédictive est alors estimée par la moyenne de ces échantillons :

$$p(y^* \mid \mathcal{D}, x^*) \approx \frac{1}{T} \sum_{t=1}^{T} p\big(y^* \mid x^*, \omega_t\big).$$

La variance de ces T prédictions fournit une estimation de l'incertitude épistémique, tandis que les sorties $p(y^* \mid x^*, \omega_t)$ pour chaque échantillon encodent l'incertitude aléatoire.

4.5 Enrichir un réseau lexico-sémantique parcimonieux

Dans les sections qui suivent, nous nous concentrons sur l'objectif d'enrichir le `RL-fr` avec de nouveaux triplets. Pour ce faire, nous nous appuyons sur les prédictions du modèle le plus performant, `CompGCN-ConvE`. Toutes les expériences sont donc appliquées aux prédictions de ce modèle.

Par ailleurs, nous appliquons les mêmes expériences sur `RezoJDM16k`, bien que ce soit un réseau lexico-sémantique très dense. L'intérêt ici n'est pas d'enrichir le réseau mais d'observer les différences en fonction de la densité du réseau considéré.

4.5.1 Application du Monte-Carlo Dropout

Afin de mesurer l'incertitude du modèle et d'obtenir des prédictions robustes, nous décidons d'appliquer la technique de Monte-Carlo Dropout (Gal and Ghahramani, 2016).

Comme décrit dans la section 4.4.3, lors d'une inférence utilisant le MC Dropout, plusieurs prédictions pour la même entrée sont générées, en échantillonnant un masque *dropout* différent à chaque fois. On obtient ainsi une distribution prédictive pour un modèle et des entrées donnés, permettant de calculer des scores de confiance pour n'importe quelle prédiction.

En reprenant l'exemple donné dans la figure 4.4 du triplet incomplet `(bonnet I, syn, ?)`, étant donnée la distribution prédictive originale du modèle, nous appliquons le MC Dropout pour générer $n = 100$ nouvelles distributions pour le triplet. Ensuite, nous définissons un critère d'inclusion pour transformer les prédictions à valeur réelle en décisions binaires. Le critère ici est de vérifier si l'entité prédite figure parmi les 10 premières prédictions.

Nous calculons le score de confiance comme le rapport entre le nombre de fois où l'entité prédite apparaît dans le top 10 et le nombre de distributions prédictives n. Cela nous permet d'établir des scores de confiance pour les prédictions du modèle et de formuler des affirmations telles que « selon ce modèle, l'entité prédite `kepi` est dans le top 10 des prédictions pour l'entité source `bonnet I` et la relation `synonymie` avec 75 % de confiance ».

4.5.2 Sélection de triplets candidats

Nous cherchons à déterminer si le score de confiance permet d'identifier des triplets pertinents susceptibles d'être ajoutés aux graphes. Pour cela, nous générons toutes les combinaisons possibles d'entités et de relations pour chaque jeu de données et nous supprimons les triplets déjà présents dans les graphes originaux (le `RL-fr` complet pour `RLF27k`). Au total, 533 551 triplets ont été générés pour `RLF27k` et 1 720 454 pour `RezoJDM16k`.

Le but étant de détecter de nouveaux potentiels triplets dans le `RL-fr`, nous nous intéressons à des triplets qui ne sont pas encore formés. Nous décidons donc d'extraire les triplets dont les entités qui ne sont pas reliées par un chemin orienté. Sur la figure 4.5, les nœuds X et Y ne sont pas reliés mais pourraient présenter un lien direct. Pour `RLF27k`, nous obtenons au total 95 766 triplets candidats.

Pour `RezoJDM16k`, en raison de la forte densité du graphe, toutes les entités sont reliées par un chemin au maximum de longueur 4. Nous ne pouvons donc pas extraire des candidats de la

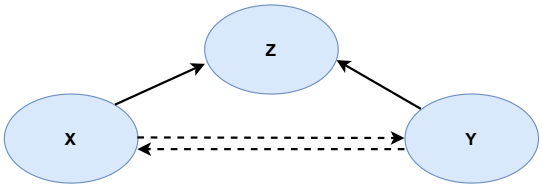


FIGURE 4.5 – Exemple d'un type de triplet candidat dans `RLF27k`. Les flèches pleines représentent les relations existantes et celles en pointillé, les relations potentielles.

même manière que pour `RLF27k`.

Nous formulons une autre hypothèse qui suppose que deux entités éloignées par un chemin de longueur maximale (4) peuvent potentiellement être reliées par une relation directe. Sur la figure 4.6, les nœuds A et D sont liés par un chemin de longueur 3. On suppose qu'il pourrait exister un triplet `(A, relation, D)`.

Nous conservons donc uniquement les triplets dont les entités sont liées par des chemins les plus longs, i.e, de longueurs 3 et 4 (le nombre de triplets de chemin de longueur 4 étant insuffisant), ce qui donne un total de 154 168 triplets candidats.

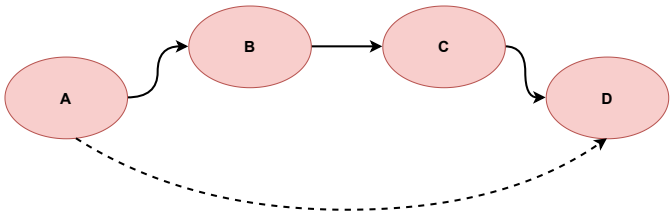


FIGURE 4.6 – Exemple d'un type de triplet candidat dans `RezoJDM16k`. Les flèches pleines représentent les relations existantes et celle en pointillé, une relation potentielle.

4.5.3 Annotations manuelles

Pour évaluer le score de confiance, 240 triplets par jeu de données ont été annotés par quatre annotateurs, chacun annotant 120 triplets. Cela nous permet ainsi d'obtenir deux annotations pour chaque triplet et de calculer un accord inter-annotateurs.

Les triplets à annoter ont des scores de confiance variés et sont représentés de manière homogène dans les échantillons de chaque annotateur, regroupés par intervalle de 0,1. La tâche d'annotation consiste à déterminer si un lien sémantique ou syntaxique existe entre deux entités. Le but n'est pas de déterminer si la relation entre les entités est correcte. Les annotateurs ne disposent d'aucune information en dehors du triplet et trois étiquettes d'annotation leur sont proposées :

— (1) il existe un lien entre les entités,

— (-1) il n'y a pas de lien,

— (0) le lien est ambigu ou discutable.

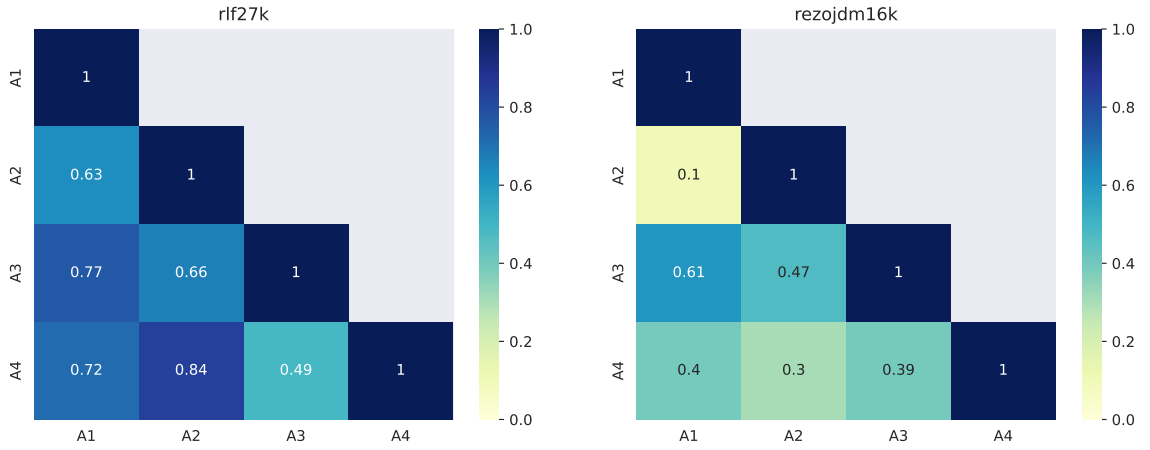


FIGURE 4.7 – Kappa de Cohen sur `RLF27k` et `RezoJDM16k`.

L'accord inter-annotateurs est calculé avec un kappa de Cohen et s'avère bien plus élevé pour le `RLF27k` avec l'accord le plus fort à 0,84 contre 0,61 pour `RezoJDM16k` (cf. Figure 4.7).

L'accord minimum pour `RezoJDM16k` est particulièrement faible, à 0,1, contrairement à 0,49 pour le `RLF27k`. La différence d'accords inter-annotateurs entre `RezoJDM16k` et `RLF27k` peut être attribuée au choix des triplets à annoter pour `RezoJDM16k` et à la distribution inégale des étiquettes d'annotation, notamment à la surreprésentation de triplets dont les entités n'ont pas de lien (étiquette 1).

En effet, sur les 183 triplets de `RezoJDM16k` où les annotations sont en accord, 85 % (156 triplets) ont été annotés comme n'ayant aucun lien entre les entités et 15 % (27) comme ayant un lien. Lors de la sélection de triplets candidats pour `RezoJDM16k` (cf. Section 4.5.2), nous avons choisi des triplets dont les entités sont éloignées par un chemin de longueur maximale (3 ou 4) en supposant qu'il peut exister un chemin de longueur 1, un chemin de longueur 3 ou 4 étant relativement court en comparaison à d'autres graphes. Cependant, en raison de sa forte connectivité, le diamètre de `RezoJDM16k` est de 4, ce qui explique pourquoi des nœuds liés par un chemin de longueur 4 sont en réalité sémantiquement distants.

La figure 4.8 présente la comparaison entre les annotations manuelles et les scores de confiance des triplets des échantillons de `RezoJDM16k` et `RLF27k`.

Pour `RezoJDM16k`, les rares triplets annotés comme exacts ont tendance à présenter des scores de confiance élevés. Toutefois, en raison de la prévalence des annotations -1, il est difficile d'établir une corrélation solide.

Concernant `RLF27k`, sur les 200 triplets où les deux annotateurs présentent un accord :

— 56,5 % (113) sont annotés comme ne présentant pas de lien (-1),

— 39,0 % (78) comme présentant un lien (1),

— 4,5 % (9) comme étant ambigus (0)

Nous pouvons noter que les étiquettes d'annotation dans `RLF27k` présentent une distribution plus équilibrée par rapport à `RezoJDM16k` et qu'on observe une corrélation plus marquée entre les

étiquettes d'annotation et les scores de confiance sur les triplets de `RLF27k`. Les triplets annotés comme -1 ont un score de confiance faible, tandis que ceux annotés comme 1 ont un score de confiance plus élevé.

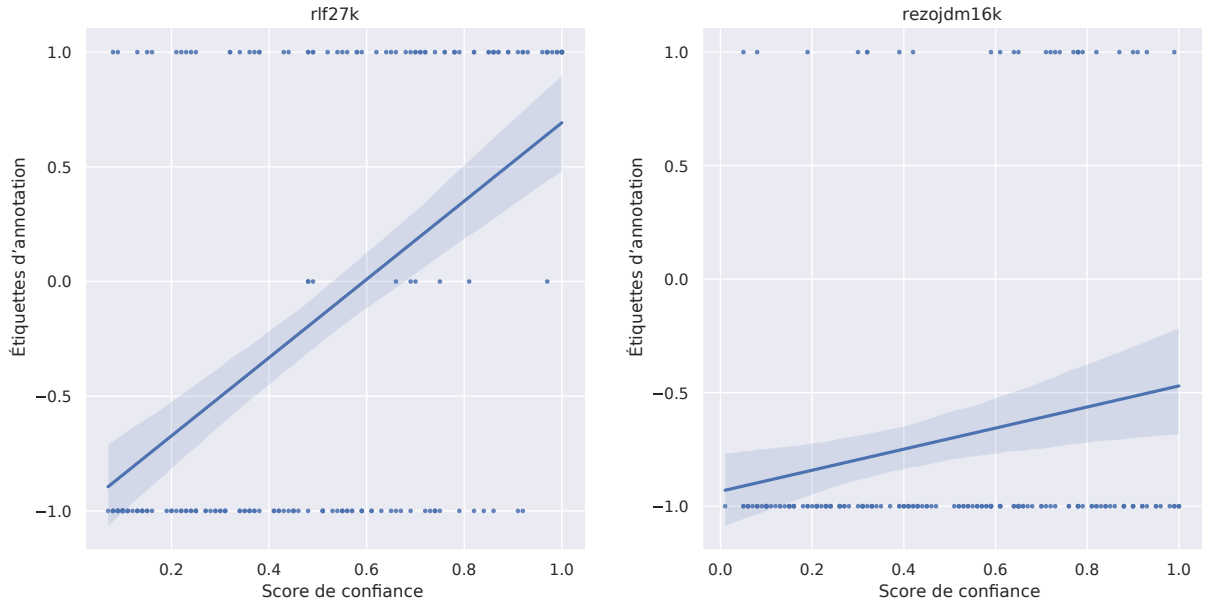


FIGURE 4.8 – Corrélations entre les valeurs d'annotation et les scores de confiance dans `RLF27k` et de `RezoJDM16k`. Les triplets considérés sont ceux pour lesquels les annotateurs sont d'accord.

4.5.4 Extraction de nouveaux triplets

Afin de déterminer un seuil de confiance pour maximiser l'exactitude des triplets à extraire, nous calculons le ratio de triplets corrects sur notre échantillon, triplets annotés par les annotateurs comme ayant un lien entre ses entités, en fonction du score de confiance (cf. Figure 4.9). Ainsi, on observe qu'un seuil de confiance élevé conduit à une plus grande proportion de triplets corrects. Par exemple, poser un seuil de confiance à 0,95 permet d'obtenir tous les triplets prédits comme étant corrects.

Dans le réseau `RL-fr`, il existe 95 766 triplets dont les entités ne sont pas connectées par un chemin orienté. En posant un seuil de confiance à 0,95, 398 triplets sont extraits et peuvent donc constituer de potentiels nouveaux triplets. Sur les 398 triplets, 43 % des triplets ont une relation de type synonymie, ce qui est cohérent avec le fait qu'il s'agit de la relation la plus présente dans le `RL-fr`. Le tableau 4.6 présente plusieurs exemples pertinents observés manuellement.

Une des limites de cette expérience réside dans le niveau de finesse de la polysémie dans le `RLF27k`. En effet, une grande partie des triplets candidats sont des triplets où le modèle associe une autre acception du vocable (cf. Tableau 4.7). Nous supposons que ce phénomène est en partie dû aux relations de copolysémie qui relient des acceptions d'un vocable. Il serait alors intéressant d'observer le changement dans les triplets candidats en supprimant ces relations du réseau.

Nous notons toutefois que le modèle prédit des entités cibles avec une étiquette de partie du discours correcte. Par exemple, la relation V_0 code la verbalisation sous la forme :

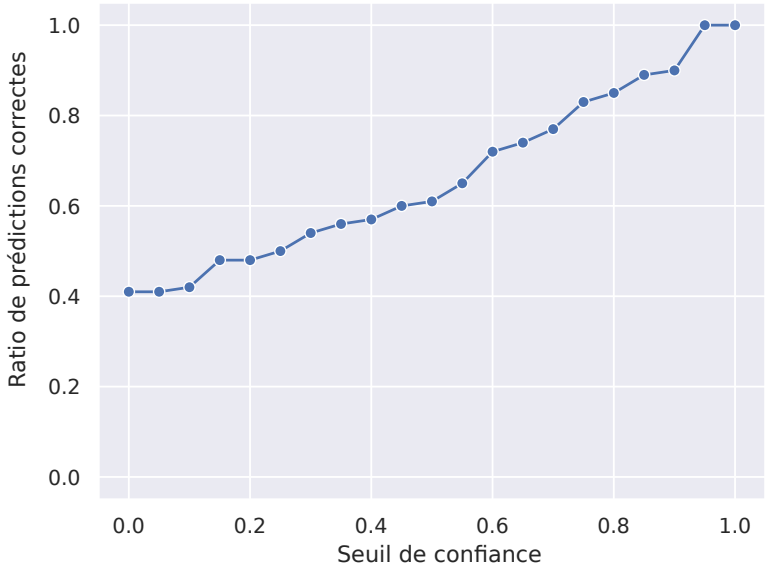


FIGURE 4.9 – Ratio de prédictions correctes (annotées 1) du `RLF27k` selon le seuil de confiance.

Triplet candidat	Score de confiance
(kidnappeur, Syn, ravisseur I)	99 %
(en aucune manière, Syn, en rien)	100 %
(motocycliste n-fem, Syn, motarde)	96 %
(marchande, Syn, débitante)	98 %

TABLE 4.6 – Exemples de triplets candidats de `RLF27k` composés d'une autre acception d'un triplet existant.

Triplet existant	Triplet candidat
(sensibilité II.2, A_1, sensible II.2)	(sensibilité II.2, A_1, sensible III)
(règne II.2, V_0, régner II.2)	(règne II.2, V_0, régner II.1)

TABLE 4.7 – Exemples de triplets candidats de `RLF27k` composés d'une autre acception d'un triplet existant.

V_0(présentation) = présenter. Sur les 25 triplets impliquant une relation de verbalisation, 24 triplets respectent la contrainte de la fonction lexicale.

4.6 Conclusion

Dans ce chapitre, nous nous inscrivons dans la continuité des travaux de Mirzapour et al. (2022) en comparant dans un premier temps, les performances des modèles de prédiction de liens sur `RezoJDM16k`, un réseau très dense et sur `RLF27k`, un réseau lexico-sémantique davantage parcimonieux en raison de sa construction manuelle par des lexicographes. Pour ce faire, nous

avons évalué les six modèles de prédiction de liens testés par Mirzapour et al. (2022) et nous avons ajouté `CompGCN-ConvE`, un modèle d'une architecture plus récente et donnant les meilleurs résultats.

Dans un second temps, nous avons exploré les prédictions finales de `CompGCN-ConvE` afin d'extraire de potentiels nouveaux triplets pour compléter `RLF27k` et ainsi améliorer sa couverture. Pour ce faire, nous avons généré des prédictions robustes en utilisant la technique de Monte Carlo Dropout, nous permettant d'obtenir un score de confiance que nous avons évalué à l'aide d'annotations manuelles. Cette phase d'annotation a démontré pour le `RLF27k` une corrélation entre les triplets avec un score de confiance élevé et la justesse du triplet déterminée par les annotateurs. Cela rend ainsi possible l'extraction de triplets pertinents à partir d'un seuil de confiance et une potentielle intégration à la ressource, sous réserve d'une validation par des experts.

Si cet ajout de triplets est profitable pour `RLF27k`, il n'apporte pas d'avantages particuliers à `RezoJDM16k`, dans la mesure où le réseau présente déjà une très forte densité. En effet, même des entités présentant une faible, voire une absence de proximité sémantique et/ou syntaxique, sont connectées par des chemins relativement courts (le diamètre du graphe étant de 4). Toutefois, nous remarquons des résultats intéressants concernant l'affinage de triplets liés par la relation générique `associated`, qui représente 31 % du réseau (cf. Tableau 4.8).

Triplet existant	**Triplet affiné**	**Score de confiance**
(infirmière, associated, personne)	(infirmière, is_a, personne)	100 %
(herpès, associated, médecine)	(herpès, domain, médecine)	100 %
(ouvrir, associated, fermer)	(ouvrir, antonym, fermer)	99 %

TABLE 4.8 – Exemples de triplets de `RezoJDM16k` avec une relation affinée.

Pour convertir les prédictions à valeur réelle en une décision binaire, nous avons retenu comme critère la présence de l'entité prédite parmi les x meilleures prédictions. Dans nos expériences, nous avons fixé arbitrairement $n = 100$ distributions de probabilités et $x = 10$ comme seuil d'inclusion. Il aurait toutefois été pertinent d'examiner l'impact de la variation de ces deux paramètres, notamment en abaissant le seuil d'inclusion, afin d'obtenir des scores de confiance plus stricts.

Par ailleurs, une deuxième limite de notre approche réside durant la phase de *negative sampling*, qui consiste à générer des triplets corrompus. En effet, ces triplets corrompus générés automatiquement par le modèle peuvent être potentiellement des triplets pertinents. Durant l'entraînement, le modèle attribue une probabilité faible aux triplets corrompus. Par conséquent, il est peu probable que ces triplets apparaissent dans le haut du classement même après application du Monte Carlo Dropout. La difficulté des graphes de connaissances est qu'une information non encodée dans le graphe peut être une information fausse comme une information vraie mais manquante. C'est pourquoi la spécificité de `RezoJDM` qui représentent des triplets faux à travers des poids négatifs est intéressante. Il aurait pu être possible d'exploiter ces triplets négatifs pour la création de triplets corrompus.

Bien que les résultats de notre approche soient prometteurs, nous maintenons que la vérification manuelle est une étape importante, car la représentation de la polysémie dans des nœuds distincts affecte directement les prédictions. Étant donné que le modèle s'appuie uniquement

sur la structure du graphe et les nœuds voisins pour saisir la sémantique, il existe des limites inhérentes à la prédiction de l'entité précise parmi les différents sens d'un mot.

Pour conclure, notre approche se distingue des précédents travaux sur la prédiction dans la mesure où nous adoptons un point de vue orienté vers l'enrichissement des ressources et non vers la performance des systèmes. Notre objectif n'est pas d'évaluer à quel point un système peut inférer des triplets artificiellement retirés d'un graphe de connaissance existant mais d'examiner les sorties concrètes du système pour aboutir à un réel enrichissement de la ressource.

Troisième partie

Évaluer l'intégration de plongements de graphe pour la désambiguïsation lexicale

Introduction

Si la deuxième partie de cette thèse explore l'apport des techniques de TAL à l'enrichissement d'un réseau lexico-sémantique, la troisième partie adopte une démarche complémentaire et inverse : elle examine comment l'intégration de ressources lexicales peut, à son tour, contribuer à l'amélioration d'un modèle de TAL, en particulier dans le cadre d'une tâche de désambiguïsation lexicale. Plus spécifiquement, nous nous intéressons à l'intégration de connaissances relatives à un réseau lexico-sémantique, sous forme de plongements de graphe, à un modèle de désambiguïsation lexicale du français.

Dans cette partie, nous traitons des questions de recherche 4 et 5 :

— QR 4 : l'intégration de plongements de graphe, appris à partir d'un réseau lexico-sémantique, peut-elle améliorer les performances d'un modèle de désambiguïsation lexicale du français ?

— QR 5 : le cadre d'évaluation de la tâche de désambiguïsation lexicale reflète-elle les performances réelles des modèles ?

Dans un premier temps, le chapitre 5 dresse un état de l'art des travaux consacrés à la désambiguïsation lexicale, l'une des tâches fondamentales de TAL. La tâche a fait l'objet de nombreux travaux, notamment grâce aux campagnes d'évaluation Senseval et SemEval qui ont permis la création de jeux de données de références et la constante création de systèmes, améliorant de plus en plus les performances.

Bien que ces performances semblent atteindre des scores proches des performances humaines, nous montrons dans un second temps que le cadre d'évaluation s'est ancré depuis des années, avec des jeux de données et des métriques qui ne reflètent pas les réelles performances des systèmes. Nous consacrons ainsi une partie importante du chapitre 5 à la remise en question de ce cadre d'évaluation et à la mise en valeur des différents biais existants.

Le chapitre 6 présente ensuite les expériences de désambiguïsation lexicale sur des données lexicographiques du français. Une différence majeure est l'utilisation d'un corpus atypique pour la tâche, du fait de ses annotations avec un inventaire de sens de granularité fine rarement utilisé, le `RL-fr`, mais également en raison de sa distribution des sens qui ne reflète pas réellement la distribution des sens d'un texte naturel.

Nous testons différents types de systèmes de désambiguïsation, allant de *baselines* simples à des grands modèles de langue auto-régressifs, très présents dans le TAL actuel. Nous exploitons également un modèle permettant d'intégrer des connaissances relatives au `RL-fr`, dont des plongements de graphe.

Au-delà de comparer les performances des systèmes, nous accordons une attention particulière à l'évaluation, avec l'utilisation de différentes métriques et l'analyse qualitative des résultats. Par ces travaux, nous cherchons à montrer que la performance d'un système ne se réduit pas à une simple valeur numérique qui peut être biaisée tant par les données que les modèles.

Chapitre 5

Désambiguïsation lexicale : état de l'art et biais d'évaluation

Sommaire

5.1 Contexte et description 95
5.2 De Senseval à SemEval : campagnes d'évaluation pour la désambiguisation lexicale 97
5.2.1 Senseval 97
5.2.2 SemEval 98
5.3 Données annotées 99
5.3.1 Annotations en sens WordNet/BabelNet 100
5.3.2 Annotations avec d'autres inventaires de sens 103
5.4 Approches principales 104
5.4.1 Approches basées sur les connaissances 104
5.4.2 Approches neuronales supervisées 105
5.4.3 Approches avec les grands modèles de langue auto-régressifs 108
5.5 Métriques traditionnelles 110
5.6 Une évaluation biaisée par un cadre avantageux 111
5.6.1 Biais de la composition des données d'évaluation 112
5.6.2 Biais de la métrique 115
5.6.3 Biais du sens le plus fréquent et des données d'entraînement 116
5.6.4 Des performances dépendantes des données 117
5.7 Conclusion 120
5.7.1 Les limites de la tâche de classification 120
5.7.2 Un cadre d'évaluation « trop » ancré 121

5.1 Contexte et description

La langue est intrinsèquement ambiguë. Plus spécifiquement, les lexèmes sont porteurs d'ambiguïté que nous appellerons ambiguïté lexicale. En linguistique, Saussure définit le signe linguistique comme une entité à deux faces complémentaires : le signifiant, la représentation graphique d'un lexème et le signifié, son concept ou sa représentation mentale (Saussure et al., 1995).

La relation entre ces deux composantes étant arbitraire mais conventionnelle, il n'existe aucune contrainte empêchant un même signifiant d'être associé à différents signifiés, ce qui se traduit

par le phénomène de polysémie : un même lexème peut présenter différents sens. Par exemple, le nom commun « avocat » présente quatre sens dans le Wiktionnaire [44] :

1.1 (Métier) (Justice) Professionnel du droit, inscrit à un barreau, chargé de défendre, d'assister une partie dans un procès et de plaider devant un tribunal.

1.2 (Par extension) Celui qui intercède pour un autre, qui la soutient, qui en défend les intérêts auprès de quelqu'un.

2. (Botanique) Fruit de l'avocatier.

3. (Colorimétrie) Couleur vert sombre soutenu tirant sur le jaune.

Si les humains sont capables de distinguer le sens d'un lexème selon le contexte, l'opération n'est pas triviale pour un système automatique qui doit encoder le contexte d'apparition de telle sorte à prédire correctement le sens du lexème. La problématique de l'ambiguïté lexicale s'est posée dans un premier temps en traduction automatique (Weaver, 1955). La résolution de la polysémie est peu à peu devenue une tâche majeure en traitement automatique des langues, appelée désambiguïsation lexicale ou *Word Sense Disambiguation* (WSD), notamment avec les campagnes d'évaluation Senseval et SemEval (cf. Section 5.2).

Formellement, on peut définir la tâche comme suit : soient un contexte C représenté par une séquence de mots-formes w_i : $C = (w_1, \ldots, w_n)$ et un vocabulaire $\mathcal{V} = \{s_i\}_i$ correspondant à l'ensemble des sens possibles des lexèmes de chaque mot-forme.

Un lexème w_i possède un ensemble de sens possibles $Senses_p(w_i) \subset \mathcal{V}$. Dans un contexte C donné tel que $w_i \in C$, w_i possède également un ensemble de sens *corrects* $Senses_c(w_i, C) \subset Senses_p(w_i)$. Bien que w_i présente généralement un unique sens correct ($|Senses_c(w_i, C)| = 1$), certains lexèmes peuvent admettre plusieurs sens corrects lorsqu'ils sont annotés par un inventaire de sens de granularité fine tel que `WordNet` (cf. Section 5.3).

L'objectif de la désambiguïsation lexicale est de construire un modèle $\mathcal{M}$ qui prédit un sens correct d'un lexème w_i étant donné un contexte C :

$$\mathcal{M}(w_i, C) \in Senses_c(w_i, C)$$

La tâche prend ainsi la forme d'une tâche de classification, où les classes correspondent aux sens possibles du lexème à désambiguïser. Contrairement à d'autres tâches de classification telles que l'étiquetage morpho-syntaxique ou la reconnaissance d'entités nommées, où l'ensemble de classes est prédéfini, en désambiguïsation lexicale, l'ensemble de classes change selon le lexème. Les lexèmes porteurs d'ambiguïté sont majoritairement les lexèmes lexicaux (ou lexèmes pleins) pouvant être des noms, des verbes, des adjectifs et des adverbes. Le nombre de sens possibles détermine alors le degré d'ambiguïté d'un lexème : plus il a de sens différents, plus il est ambigu, plus la tâche de désambiguïsation est difficile.

Dans ce chapitre, nous présentons dans un premier temps un état de l'art de la tâche de désambiguïsation lexicale. En section 5.2, nous commençons par donner un historique des campagnes d'évaluation Senseval et SemEval, qui ont joué un rôle majeur dans le développement de méthodes et de données pour la tâche. Dans la section 5.3 sur les jeux de données, nous présentons des *frameworks* classiques et des données plus spécifiques pour le français. Les approches

44. `https://fr.wiktionary.org/wiki/avocat`, consulté le 22 octobre 2025.

sont ensuite détaillées dans la section 5.4 avec une classification en trois types : les approches basées sur les connaissances, les approches neuronales supervisées et les approches utilisant les grands modèles de langue. Nous détaillons également les métriques traditionnellement utilisées pour évaluer les systèmes.

Dans un second temps, nous portons une attention particulière à l'évaluation, qui s'est ancrée dans un schéma fixe, entraînant une course à la performance au détriment d'une analyse qualitative des résultats. Nous consacrons ainsi la section 5.6 à une analyse des métriques et des biais liés aux données.

Enfin, en conclusion, nous discutons dans la section 5.7.1 des variations de la tâche de désambiguïsation lexicale qui découlent des limites de sa configuration traditionnelle en tâche de classification et de l'évolution des systèmes automatiques de TAL.

5.2 De Senseval à SemEval : campagnes d'évaluation pour la désambiguisation lexicale

En TAL, la tâche de désambiguïsation lexicale a fait l'objet de nombreux travaux, notamment dans le cadre des campagnes d'évaluation ou *shared tasks* Senseval, de 1998 à 2004, puis SemEval à partir de 2007.

SemEval est une série d'ateliers internationaux visant à faire progresser la recherche sur l'analyse sémantique et la création de données annotées pour la résolution de différentes tâches sémantiques [45]. L'atelier propose une série de campagnes d'évaluation où différentes équipes présentent leurs systèmes. La tâche de désambiguïsation lexicale a été une tâche majeure de ces ateliers, permettant l'élaboration continue de systèmes et la création de jeux de données devenus des références pour l'évaluation.

5.2.1 Senseval

En 1998, **Senseval-1** devient la première campagne d'évaluation de la désambiguïsation lexicale sur trois langues : l'anglais, le français et l'italien. Pour l'anglais, les deux corpus à disposition sont le `SemCor` (Miller et al., 1994) et `DSO` (Ng and Lee, 1996), manuellement annotés en sens `WordNet`. Les approches proposées sont majoritairement basées sur des règles et des définitions dictionnairiques (Kilgarriff and Palmer, 1998).

En 2001, la deuxième édition, **Senseval-2**, vise à inclure davantage de langues pour trois sous-tâches de la désambiguïsation lexicale (Edmonds and Cotton, 2001) :

— la tâche *all-words* où les systèmes doivent étiqueter en sens tous les lexèmes pleins d'un échantillon de texte (en tchèque, néerlandais, anglais et estonien),

— la tâche *lexical sample* où un échantillon de lexèmes défini doit être étiqueté dans différents textes (en basque, anglais, italien, japonais, coréen, espagnol et suédois),

— la tâche de traduction, uniquement sur du japonais vers l'anglais où le sens est décrit par une traduction en anglais.

Si le corpus `SemCor` et l'inventaire de sens `WordNet` sont devenus les jeux de données de référence pour l'anglais, les équipes ont pu proposer leurs propres corpus et inventaires de sens

45. `https://semeval.github.io/`

pour les autres langues. Les versions italienne, espagnole et estonienne du `WordNet` créées dans le cadre du projet `EuroWordNet` (Vossen, 1998) ont par exemple été exploitées. Les systèmes proposés dans Senseval-2 sont davantage basés sur des méthodes d'apprentissage machine utilisant des classifieurs naïfs bayésiens, des arbres de décision ou la méthode des k plus proches voisins (Yarowsky et al., 2001; Kilgarriff, 2001).

Dans l'édition **Senseval-3** de 2004, la tâche de désambiguïsation lexicale a été étendue à de nouvelles langues comme le chinois, le roumain et le catalan. Cette édition a par ailleurs intégré une nouvelle tâche, l'identification de rôles sémantiques (Mihalcea and Edmonds, 2004). La première tâche *all-words* sur l'anglais a été une des tâches majeures avec 26 systèmes proposés par 16 équipes (Snyder and Palmer, 2004). Avec une évaluation renforcée par des jeux de données de test diversifiés et élargis, les systèmes supervisés ont présenté de meilleures performances que les systèmes non supervisés. Les résultats atteignent un plateau entre 65-70 % de précision et rappel, ce qui est cohérent avec l'accord inter-annotateurs de 70-75 %.

En termes de systèmes, cette édition marque la montée des modèles statistiques supervisés et des approches basées sur les relations sémantiques de `WordNet` (Mihalcea and Faruque, 2004; Decadt et al., 2004).

5.2.2 SemEval

Les systèmes et applications sémantiques allant au-delà de la désambiguïsation lexicale, Senseval a évolué vers SemEval, qui vise à évaluer davantage d'aspects des systèmes sémantiques computationnels en proposant une plus grande diversité de tâches.

En 2007, la tâche 17 de **SemEval-2007** traite la désambiguisation lexicale sous un aspect plus large avec l'intégration de la sous-tâche d'étiquetage des rôles sémantiques (Pradhan et al., 2007). Pour la tâche *all-words*, le système le plus performant est constitué d'un classifieur basé sur le principe d'entropie maximale avec une F1 de 59,1 % contre la *baseline* utilisant le sens le plus fréquent à 51,4 %.

À l'issue de cette tâche, les équipes ont notamment observé que les modèles présentent des performances plus élevées sur des données annotées avec un inventaire de sens « à gros grain » (*coarse-grained*). De plus, les performances tendent à montrer que la désambiguïsation des verbes est plus difficile que celle des noms, potentiellement du fait d'une grande proportion de verbes à forte ambiguïté (Pradhan et al., 2007).

Dans l'édition **SemEval-2013**, la tâche de désambiguisation lexicale (tâche 12) a été proposée dans un contexte multilingue avec des données en anglais, français, espagnol, allemand et italien (Navigli et al., 2013). L'inventaire de sens utilisé est `BabelNet` (Navigli and Ponzetto, 2012), un dictionnaire encyclopédique multilingue créé à partir de `WordNet` et `Wikipedia`.

Les systèmes sont évalués sur un échantillon de test anglais traduit dans les quatre autres langues. Seuls les noms ont été annotés en sens manuellement, `BabelNet` ne couvrant les autres classes qu'en anglais. Tous les systèmes proposés sont des systèmes basés sur le graphe et ses propriétés structurales plutôt que sur l'apprentissage supervisé. Le système le plus performant `UMCC-DLSI` adapte le modèle *Personalized PageRank* (Agirre and Soroa, 2009) et surpasse la *baseline* du sens le plus fréquent dans toutes les langues sauf en allemand, avec une F1 de 0,65-0,71 contre 0,57-0,67 pour la *baseline*.

Dans **SemEval-2015**, la tâche de désambiguïsation lexicale multilingue (tâche 13) est proposée avec la tâche de *Entity linking*. La désambiguisation lexicale nécessitant des inventaires de sens comme `WordNet` et la reconnaissance d'entités nommées, des inventaires encyclopédiques tels que `Wikipedia` ou `DBpedia`, l'objectif est de résoudre conjointement ces deux tâches en utilisant une ressource commune, `BabelNet`. Les données disponibles sont dans trois langues : l'anglais, l'espagnol et l'italien et couvrent trois domaines : biomédical (description de médicaments), mathématiques et informatique (manuels de logiciels) et social (rapports sur les politiques de l'UE). Les données sont issues de corpus parallèles et sont manuellement annotées en sens `BabelNet` par des locuteurs natifs (Moro and Navigli, 2015).

Les résultats montrent que les systèmes non supervisés basés sur les graphes et la similarité sémantique produisent de meilleures performances dans un contexte multilingue que les systèmes supervisés. Les performances sont également variables selon le domaine auquel les données sont relatives : les données biomédicales ont été les mieux prédites due à une faible polysémie (F1 = 71,3 % avec le meilleur système) contrairement aux données de mathématiques et informatique qui présentent une F1 inférieure à 55 %.

À l'issue de cette édition, les organisateurs ont proposé en perspective l'extension de la tâche à des langues non européennes et peu dotées, le développement de corpus multilingues annotés en sens et de systèmes de désambiguïsation spécifiques à des domaines (Moro and Navigli, 2015).

Bien que la tâche de désambiguïsation lexicale sous la configuration classique n'a plus été proposée après SemEval-2015, de nombreux travaux se sont intéressés au développement de systèmes que nous détaillons dans la section 5.4. Nous présentons également les limites et les variations de la tâche dans la section 5.7.1.

5.3 Données annotées

La tâche de désambiguïsation lexicale requiert deux ressources principales : un **inventaire de sens** qui décrit les différents sens d'un lexème et des **données annotées en sens** pour l'entraînement et l'évaluation. Les catégories grammaticales annotées sont généralement les noms communs, les verbes, les adjectifs et les adverbes. Les instances annotées sont toutefois variables selon les jeux de données. La figure 5.1 présente une phrase annotée en sens `WordNet` tirée du jeu de données ALL_{EN}.

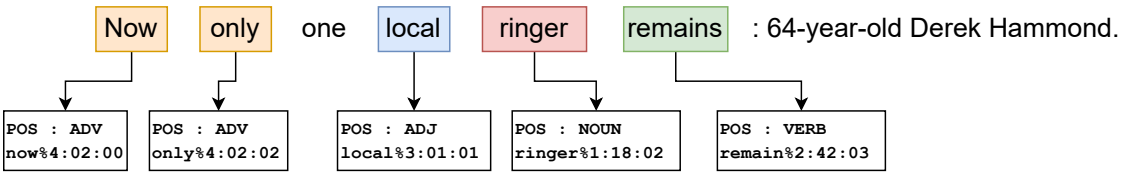


FIGURE 5.1 – Annotations d'une phrase en sens `WordNet` dans ALL_{EN} (senseval2.d000.s008).

La complexité de la désambiguïsation dépend fortement de la granularité de l'inventaire de sens. Ainsi, certaines instances peuvent admettre plusieurs sens corrects (plusieurs *gold*), que nous appellerons des instances *multi-gold*. Les données annotées en sens `WordNet` présentent certaines instances *multi-gold*, `WordNet` étant de granularité fine. Par exemple :

You can set each graph its own **color**. [46]

46. Phrase semeval2015.d001.s053 de ALL_{EN}.

L'instance **color** est annotée avec deux sens corrects, que nous décrivons avec leurs définitions respectives :

— `color%1:07:01` : the appearance of objects (or light sources) described in terms of a person's perception of their hue and lightness (or brightness) and saturation.

— `color%1:07:00` : a visual attribute of things that results from the light they emit or transmit or reflect.

Les annotations en sens peuvent ainsi varier entre les jeux de données, d'une part, en termes de nombre d'annotations car tous les lexèmes pleins ne sont pas systématiquement annotés et d'autre part, en termes de granularité, celle-ci étant variable selon l'inventaire de sens.

Dans cette section, nous présentons les données principales disponibles pour la tâche, selon l'inventaire de sens utilisé pour les annotations. Le tableau 5.1 représente une vue d'ensemble des *frameworks* et des données que nous détaillerons.

	Langues	**Inventaire de sens**	**Données d'entraînement**	**Données de test**
Framework Raganato et al. (2017)	anglais	`WordNet 3.0`	`SemCor`, `OMSTI`	ALL_{EN} : Senseval, SemEval
ALL_{NEW} (Maru et al., 2022)	anglais	`WordNet 3.0`	—	4 917 instances
`42D` (Maru et al., 2022)	anglais	`WordNet 3.0`	—	370 instances
`hardEN` (Maru et al., 2022)	anglais	`WordNet 3.0`	—	476 instances
`XL-WSD` (Pasini et al., 2021)	multilingue	`BabelNet 4.0`	`SemCor`, `WNG` (*silver-data*)	Senseval, SemEval
`FLUE` (Le et al., 2020)	français	`WordNet 3.0`	`SemCor`, `WNG` (*silver-data*)	SemEval2013-FR (adapté)
`FrenchSemEval` (Segonne et al., 2019)	français	`Wiktionary`	Exemples de `Wiktionary`	3 199 instances
Sinha et al. (2022)	français	`RL-fr`	`BEL-RL-fr`	partie du `BEL-RL-fr`

TABLE 5.1 – Tableau récapitulatif des données principales disponibles pour la désambiguïsation lexicale.

5.3.1 Annotations en sens WordNet/BabelNet

La majorité des données utilisées en désambiguisation lexicale utilise comme inventaire de sens `WordNet`, qui a été choisi dès le lancement des campagnes d'évaluation Senseval et SemEval. Le développement de `WordNets` dans d'autres langues a permis d'étendre la tâche à un contexte multilingue, d'autant plus avec l'utilisation de `BabelNet` en tant qu'inventaire de sens. Deux *frameworks* en particulier sont devenus des références pour l'évaluation de systèmes (Bevilacqua et al., 2021) : le *framework* de Raganato et al. (2017) pour l'anglais et le *framework* multilingue, `XL-WSD` (Pasini et al., 2021).

5.3.1.1 *Framework* de Raganato et al. (2017)

Les jeux de données d'évaluation provenant des multiples éditions de Senseval et SemEval n'étant pas standardisés en raison des différentes versions de `WordNet`, Raganato et al. (2017) proposent un *framework* unifié des données de l'anglais, converties aux sens de la version 3.0 de `WordNet`. Les données d'évaluation sont composées des cinq jeux de données *all-words* de granularité fine de Senseval-2, Senseval-3, SemEval-2007, SemEval-2013 et SemEval-2015. Dans cette thèse, nous désignerons ce jeu de données d'évaluation, ALL_{EN}.

Quant aux données d'entraînement, elles proviennent de deux ressources :

— `SemCor` (Miller et al., 1994), corpus contenant plus de 200 000 annotations en sens `WordNet` mais ne couvrant que 22 % des 118 000 *synsets* de `WordNet`.

— `OMSTI` (Taghipour and Ng, 2015), corpus construit par alignement automatique du corpus parallèle anglais-chinois, `MultiUN` (Eisele and Chen, 2010).

La figure 5.2 présente les statistiques relatives aux données du *framework*. Ces données ont été majoritairement utilisées pour l'évaluation et la comparaison des systèmes de désambiguïsation lexicale, particulièrement les données en anglais.

	#Docs	#Sents	#Tokens	#Annotations	#Sense types	#Word types	Ambiguity
Senseval-2	3	242	5,766	2,282	1,335	1,093	5.4
Senseval-3	3	352	5,541	1,850	1,167	977	6.8
SemEval-07	3	135	3,201	455	375	330	8.5
SemEval-13	13	306	8,391	1,644	827	751	4.9
SemEval-15	4	138	2,604	1,022	659	512	5.5
SemCor	352	37,176	802,443	226,036	33,362	22,436	6.8
OMSTI	-	813,798	30,441,386	911,134	3,730	1,149	8.9

FIGURE 5.2 – Statistiques des jeux de données de désambiguïsation lexicale des campagnes Senseval et SemEval après unification (Table 1 de Raganato et al. (2017)).

5.3.1.2 XL-WSD (Pasini et al., 2021)

`XL-WSD` est un *framework* d'évaluation pour la désambiguïsation lexicale multilingue, annoté en sens `BabelNet 4.0` (aligné avec `WordNet 3.0`) et couvrant 18 langues de six familles de langues différentes :

— langues indo-européennes : l'allemand, l'anglais, le bulgare, le catalan, le croate, le danois, l'espagnol, le français, le galicien, l'italien, le néerlandais, le slovène ;

— langue japonique : le japonais ;

— langue coréanique : le coréen ;

— langue sino-tibétaine : le chinois ;

— langues ouraliennes : le hongrois, l'estonien ;

— et une langue isolée : le basque.

Les données d'entraînement sont des *silver-data* obtenues par traduction automatique des corpus `SemCor` et `Princeton WordNet Gloss Corpus` (`WNG`) [47] disponibles en anglais. Le corpus `WNG` a été créé à partir des définitions et des exemples des *synsets* dans `WordNet`. L'annotation a été effectuée *via* un processus manuel et semi-automatique (Pasini et al., 2021). Le corpus contient au total 614 435 instances annotées avec 117 653 *synsets* différents. L'ajout de `WNG` dans les données d'entraînement permet de pallier la faible couverture de `SemCor` des *synsets* de `WordNet` (22 % des *synsets* seulement).

Quant aux données d'évaluation, elles proviennent des campagnes de SemEval pour la résolution de la tâche dans un contexte multilingue. Les données ont été annotées en sens `BabelNet` en exploitant l'alignement entre les `WordNets` respectifs à chaque langue et le `WordNet 3.0`, puis l'alignement entre `WordNet 3.0` et `BabelNet 4.0`.

47. `https://wordnetcode.princeton.edu/glosstag.shtml`

5.3.1.3 ALL_{NEW}, 42D, hardEN (Maru et al., 2022)

Maru et al. (2022) remettent en question l'évaluation de la tâche de désambiguïsation lexicale, notamment l'utilisation de la micro-F1 qui peut biaiser les performances vers les classes les plus représentées (cf. Section 5.6). Avec une analyse des instances mal prédites par la majorité des systèmes, ils démontrent que ces systèmes ont tendance à privilégier le premier sens dans WordNet [48] et ceux présents dans les données d'entraînement. Ils corrigent également les erreurs dans les instances de ALL_{EN} avec l'aide d'un linguiste et créent une version mise à jour, appelée ALL_{NEW}. Le jeu de données est constitué de 4 917 instances polysémiques provenant de ALL_{EN}. Les instances monosémiques ont été retirées et des erreurs d'annotation ont été corrigées.

De plus, afin de refléter plus efficacement les capacités des systèmes de désambiguïsation, ils introduisent deux nouveaux jeux de données en anglais, constitués d'instances complexes.

42D est composé de 370 instances de test manuellement annotées provenant du British National Corpus (Leech, 1992). Le jeu de données se veut plus exigeant que les jeux d'évaluation classique pour plusieurs raisons :

— les sens corrects ne sont pas dans SemCor,

— les sens corrects ne sont pas les sens les plus fréquents,

— il n'y a pas d'instances monosémiques,

— les instances proviennent de 42 domaines différents définis dans BabelNet 4.0.

hardEn est une intersection entre les instances de ALL_{NEW}, 42D et $S10_{NEW}$ [49] que les systèmes n'ont pas réussi à résoudre. Il contient au total 476 instances de test.

5.3.1.4 FLUE (Le et al., 2020)

Inspiré de GLUE (Wang et al., 2018), FLUE est un *benchmark* créé pour évaluer des systèmes de TAL sur des tâches syntaxiques et sémantiques du français, dont la tâche de désambiguïsation des tâches sur les noms et les verbes.

Pour les noms, les données d'entraînement sont constitués de traductions automatiques de SemCor et WordNet Gloss Corpus de l'anglais vers le français. Les annotations en sens WordNet sont ajoutées sur les lexèmes traduits à partir d'un système d'alignement. Pour les données d'évaluation, les données de test en français proposées dans la désambiguïsation lexicale multilingue (tâche 12) de SemEval 2013 sont récupérées. Originellement annotées avec des sens BabelNet, les annotations sont converties automatiquement en sens WordNet lorsqu'une projection est présente dans BabelNet. Dans le cas contraire, l'instance est retirée. Après une vérification manuelle, le corpus contient 306 phrases et 1 445 lexèmes français annotés en sens WordNet.

Pour les verbes, Le et al. (2020) reprennent les données du FrenchSemEval (Segonne et al., 2019) (cf. Section 5.3.2.1).

48. Les sens dans WordNet sont classés par fréquence d'apparition dans le corpus SemCor.

49. Version corrigée des instances de SemEval-2010 (tâche 17) (Agirre et al., 2010) proposés par Maru et al. (2022).

5.3.2 Annotations avec d'autres inventaires de sens

Bien que plus rares, des travaux se sont intéressés à la création de données annotées avec un inventaire de sens différent de WordNet ou BabelNet, pour pallier une granularité trop fine des sens, la sous-représentation de certains sens ou le manque de corpus annotés avec d'autres ressources. Dans cette sous-section, nous présentons en particulier des ressources annotées avec Wiktionary [50] et le RL-fr (Lux-Pogodalla and Polguère, 2011).

5.3.2.1 Utilisation de Wiktionary

FrenchSemEval (Segonne et al., 2019; Segonne, 2021). Pour évaluer la désambiguïsation lexicale des verbes en français, Segonne et al. (2019) créent FrenchSemEval, un jeu de données de verbes annotés avec des sens du Wiktionary français.

Bien qu'ils aient initialement utilisé Eurosense (Delli Bovi et al., 2017), un corpus multilingue annoté automatiquement en sens BabelNet, leur évaluation manuelle a montré que les annotations BabelNet étaient d'une granularité trop fine pour pouvoir désambiguïser les verbes, avec une moyenne de 15,5 sens par verbe sur un échantillon de 150 phrases. Ils remarquent également que la différence entre certains sens était difficile à juger.

Ils construisent alors FrenchSemEval, un corpus d'évaluation constitué de 3 199 occurrences de verbes manuellement annotées (environ 50 occurrences par verbe), utilisant une version du Wiktionary français disponible *via* DBnary (Sérasset, 2015). Les données d'entraînement sont composées d'exemples de Wiktionary avec au total 55 206 annotations.

FEWS (Blevins et al., 2021). Le jeu de données FEWS (Few-shot Examples of Word Senses) a été créé explicitement pour permettre la désambiguïsation lexicale *low-shot* et *zero-shot*. Il est construit à partir des exemples du Wiktionary anglais.

Blevins et al. (2021) exploitent Wiktionary pour sa grande couverture de sens, permettant d'inclure des sens rares et spécifiques à des domaines, sens non représentés dans les jeux de données annotés avec WordNet.

5.3.2.2 Utilisation du RL-fr

Le RL-fr est un réseau lexico-sémantique du français encodant des relations entre des unités lexicales à travers des fonctions lexicales (Lux-Pogodalla and Polguère, 2011). Basées sur la théorie Sens-Texte, les fonctions lexicales permettent de modéliser des liens paradigmatiques et syntagmatiques entre les unités lexicales (cf. Section 1.3.3).

Afin de compléter le RL-fr avec des exemples lexicographiques, le corpus BEL-RL-fr a été créé à partir de Frantext [51] et des articles de l'Est Républicain (Lux-Pogodalla, 2014). La version la plus récente du BEL-RL-fr est constituée de 32 284 d'exemples et de 53 633 segments textuels annotés en unité lexicale du RL-fr [52]. Sinha et al. (2022) ont exploité le corpus pour la tâche de désambiguïsation lexicale, explorant notamment l'intégration d'informations inhérentes au réseau lexico-sémantique pour l'amélioration des performances. La meilleure configuration atteint une exactitude de 56,15 % sur les noms et 38,38 % sur les verbes.

Initialement, le BEL-RL-fr a été créé par des lexicographes dans le but d'illustrer chaque sens du RL-fr avec un exemple d'usage. La motivation première n'étant pas de construire un corpus

50. https://www.wiktionary.org/
51. https://www.frantext.fr/
52. https://www.ortolang.fr/market/corpora/examples-ls-fr, consulté le 3 décembre 2025.

pour la désambiguïsation lexicale, le corpus ne reflète pas la distribution réelle des sens dans un texte naturel. Cette problématique est davantage explorée dans nos expériences détaillées dans le chapitre 6.

5.4 Approches principales

La désambiguïsation lexicale a été un enjeu majeur en TAL, entraînant le développement continu de systèmes, allant des méthodes symboliques aux techniques fondées sur l'apprentissage profond. L'évolution diachronique des systèmes est particulièrement visible dans les campagnes d'évaluation Senseval et SemEval qui ont eu lieu à partir de 1998 jusqu'en 2015 (cf. Section 5.2). Nous suivons la classification des approches de Bevilacqua et al. (2021), en y ajoutant les approches avec les grands modèles de langue :

— les approches basées sur les connaissances,

— les approches neuronales supervisées,

— les approches avec les grands modèles de langue.

5.4.1 Approches basées sur les connaissances

Les premiers systèmes de désambiguïsation lexicale sont basés sur des **méthodes symboliques**, exploitant des ressources externes telles que des dictionnaires. Une méthode connue est l'algorithme de Lesk (Lesk, 1986) qui consiste à calculer l'intersection entre la phrase dans son contexte et les définitions des sens. Bien que la méthode soit relativement simple, elle ne passe pas bien à l'échelle car le calcul de l'intersection est proportionnel à la taille de toutes les définitions.

Banerjee and Pedersen (2002) modifient l'algorithme Lesk pour exploiter les relations entre les synonymes dans le `WordNet`. Contrairement à l'approche originale qui utilise uniquement les définitions des sens possibles, ils comparent les définitions des lexèmes reliés au lexème à désambiguïser. Cela permet de fournir davantage d'informations et de contexte. Leurs résultats sur les données Senseval-2 montrent une exactitude de 32 % contre 23 % avec l'algorithme original.

En testant des variantes de l'algorithme, Vasilescu et al. (2004) montrent que les définitions et les exemples ne suffisent pas pour la désambiguïsation et qu'il est nécessaire d'intégrer des informations syntaxiques ou pragmatiques pour atteindre de meilleures performances.

La majorité des données pour la désambiguïsation étant annotées avec `WordNet`, plusieurs approches se sont appuyées sur **la structure du graphe** et sa connectivité, en exploitant les marches aléatoires (Agirre and Soroa, 2009; Moro et al., 2014). Un exemple notable est `SyntagRank` (Scozzafava et al., 2020), le système le plus performant pour la désambiguïsation multilingue et qui applique l'algorithme Personalized Page Rank (Page et al., 1999) sur la portion `WordNet` de `BabelNet` et sur `SyntagNet` (Maru et al., 2019), une ressource qui présente des relations entre les *synsets* dont les sens forment une collocation.

Les définitions ont également été exploitées d'une autre manière avec $SREF_{KB}$ (Wang and Wang, 2020). L'approche se base sur des plongements de sens générés avec `BERT` (Devlin et al., 2019) à partir d'exemples et de définitions dans `WordNet`. Chaque plongement de sens est ensuite amélioré à partir des relations entre les *synsets*. Testé uniquement sur ALL_{EN}, les données de

test anglais de Raganato et al. (2017), le système atteint une F1 de 73,5 % contre 71,7 % pour `SyntagRank`.

Les méthodes basées sur des connaissances, que ce soit des dictionnaires, des définitions ou des informations structurelles d'un graphe sémantique, ont pour avantage de ne pas nécessiter de grandes quantités de données annotées. Elles sont donc applicables à des langues peu dotées. Toutefois, les performances de ces systèmes sont fortement dépendantes de la qualité et de la richesse des ressources, qui représentent les informations de manière statique rendant difficile la prise en compte du contexte.

5.4.2 Approches neuronales supervisées

Avec l'expansion des modèles de langue basés sur l'architecture *transformer* (Vaswani et al., 2017), les systèmes neuronaux se sont largement développés grâce à leurs capacités à encoder du texte et son contexte.

La plupart des modèles supervisés de désambiguïsation lexicale traite la tâche comme une tâche de classification contextuelle. En pratique, les modèles sont entraînés dans le but de calculer, pour un lexème w dans un contexte donné C, une distribution de probabilités P_w sur tous les sens s du vocabulaire, $s \in V$. Étant donnée une phrase représentée par une séquence de tokens, un modèle de langue pré-entraîné tel que `BERT` génère un plongement contextuel pour chaque token. Ces plongements sont ensuite donnés en entrée à un classifieur entraîné sur la tâche de désambiguïsation, qui produit une distribution de probabilités sur tous les sens du vocabulaire. Le sens ayant la probabilité maximale parmi les sens possibles du lexème-cible constitue la prédiction du modèle (cf. Figure 5.3).

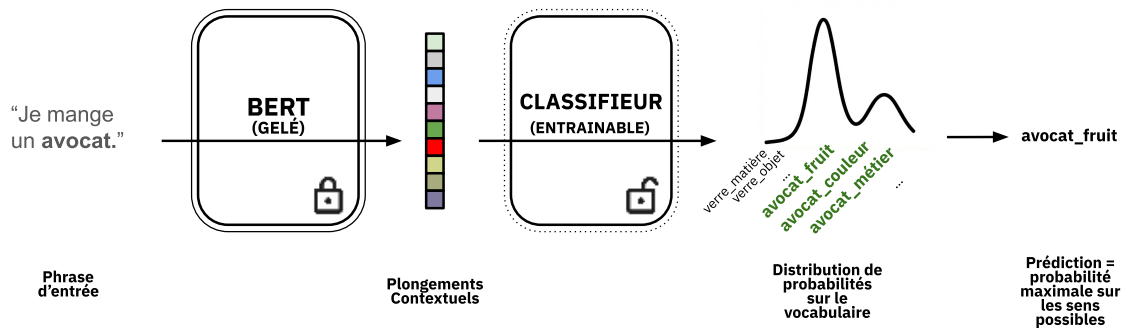


FIGURE 5.3 – *Pipeline* classique d'un modèle neural supervisé de désambiguïsation lexicale.

Les variations des modèles supervisés neuronaux ont généralement lieu au niveau du classifieur, dont l'architecture peut changer et prendre la forme d'un perceptron multicouche (Hadiwinoto et al., 2019) ou d'un ensemble de couches de *transformer* (Bevilacqua and Navigli, 2019; Vial et al., 2019).

Afin d'améliorer les systèmes supervisés, un grand nombre de travaux se sont intéressés à l'ajout de connaissances, généralement celles utilisées dans les premières approches symboliques et basées sur les graphes. Dès les approches non supervisées, les définitions ont constitué des

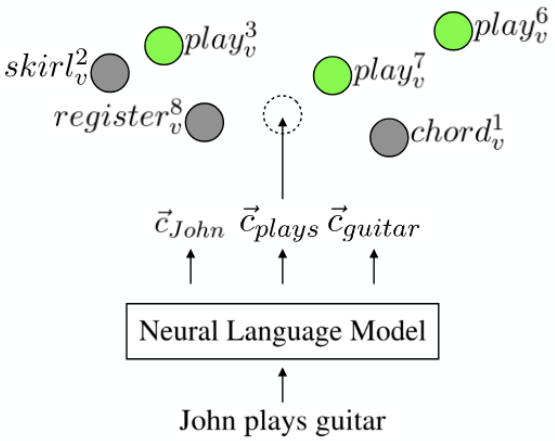


FIGURE 5.4 – Méthode des k plus proches voisins pour la désambiguïsation lexicale. Les plongements de sens et les plongements contextuels de la phrase d'entrée sont dans le même espace vectoriel. (Figure 1 de Loureiro and Jorge (2019)).

informations exploitables pour la désambiguïsation dans la mesure où elles fournissent une description des différents sens d'un lexème, sous un format textuel. En outre, un graphe sémantique contient également des informations sur les relations entre les sens, représentables sous la forme de matrices.

Les modèles supervisés peuvent être alimentés par ces informations supplémentaires, notamment au niveau du classifieur, afin d'affiner les représentations contextuelles des sens. Nous présentons ici les principaux modèles supervisés, en détaillant la manière dont ces informations sont intégrées.

Dans leurs travaux, Loureiro and Jorge (2019) présentent `LMMS`, une méthode qui résout la désambiguïsation lexicale avec la **méthode des k plus proches voisins**. L'approche consiste à générer des plongements de sens pour tous les sens dans `WordNet` avec des modèles de langue pré-entraînés (`BERT` et `ELMo`). Ces plongements sont produits à partir de `SemCor` et pour les sens non présents dans le corpus, le modèle exploite la structure sémantique de `WordNet` pour obtenir un plongement de sens à partir des sens voisins. Des plongements statiques et des définitions sont également utilisés pour enrichir ces plongements de sens.

Pour résoudre la désambiguïsation lexicale, le modèle reçoit une phrase en entrée et génère des plongements contextuels. Les plongements contextuels et les plongements de sens étant dans le même espace vectoriel, il est possible de calculer une similarité cosinus avec les k plus proches voisins (cf. Figure 5.4). Leur meilleure configuration fait état d'une micro-F1 de 75,4 % sur ALL_{EN}.

Huang et al. (2019) proposent une autre approche avec `GlossBERT`, qui consiste à reformuler la tâche de désambiguïsation lexicale en une **tâche de classification de paire phrase-définition**. En pratique, pour un sens candidat, la phrase d'entrée et la définition du sens sont concaténées en une séquence, qui est donnée à un modèle `BERT`. Il calcule ensuite un score pour chaque paire et récupère celle avec le score le plus élevé. Avec leur meilleure configuration, ils obtiennent une micro-F1 de 77,0 % sur ALL_{EN}. Si le principe est relativement simple à appliquer et permet de traiter des sens de granularité fine, le coût en calcul est élevé en raison de la combinaison de la phrase avec chaque sens.

En plus de proposer $SREF_{KB}$ (cf. Section 5.4.1), Wang and Wang (2020) construisent également une variation appelée $SREF_{sup}$ basée sur des plongements de sens supervisés obtenus par apprentissage sur `SemCor`. Ils combinent l'approche de Loureiro and Jorge (2019) et leurs plongements de sens améliorés avec les relations d'hyperonymie et d'hyponymie des *synsets*. Le modèle $SREF_{sup}$ obtient de meilleurs résultats que $SREF_{KB}$ puisqu'il produit une micro-F1 de 77,8 % contre 73,5 % pour $SREF_{KB}$.

Un autre système exploitant les définitions est `BEM` (Bi-Encoder Model) (Blevins and Zettlemoyer, 2020), **basé sur deux encodeurs** différents, un premier pour générer des plongements de sens à partir de définitions et un deuxième pour apprendre le contexte. À l'inférence, une similarité cosinus est calculée entre le plongement du contexte et les plongements de sens. Blevins and Zettlemoyer (2020) montrent que l'architecture bi-encodeurs permet de mieux généraliser pour les sens peu fréquents et en *zero-shot*, avec une micro-F1 finale de 79 % sur ALL_{EN}.

Les systèmes supervisés étant biaisés vers le sens le plus fréquent, en particulier pour les lexèmes non vus durant la phase d'entraînement (cf. Section 5.6), Kumar et al. (2019) proposent `EWISE` (Extended WSD Incorporating Sense Embeddings), un **modèle supervisé combinant des données annotées en sens, des définitions et informations provenant du graphe**. Les plongements de sens sont générés avec un encodeur à partir des définitions et de plongements de graphe. L'approche consiste alors à utiliser ces plongements de sens pour les prédictions, au lieu d'étiquettes de sens discrètes. Le fait de prédire dans l'espace continu des plongements de sens permet de généraliser la prédiction en *zero-shot* des sens inconnus.

Bevilacqua and Navigli (2020) dérive `EWISE` avec `EWISER`, un modèle supervisé neuronal qui intègre les informations relationnelles de `WordNet` directement dans l'architecture du modèle, à travers une **matrice d'adjacence encodant les poids des relations entre sens** (cf. Section 6.3.2). Leur approche exploite également des plongements de sens obtenus à partir des définitions. Le modèle obtient une micro-F1 de 80,1 % dans les résultats d'origine.

Pour améliorer la capacité du modèle à généraliser, Barba et al. (2021a) propose `ESCHER`, une autre approche qui reformule la tâche de désambiguïsation lexicale en une **tâche d'extraction de séquences** (*span extraction*), permettant au modèle d'observer conjointement le contexte de la phrase et les définitions des sens. Le principe consiste à donner en entrée du modèle la phrase à désambiguïser et la concaténation des définitions possibles. Le modèle doit ensuite extraire la séquence de texte dans cette concaténation de définitions qui correspond au sens approprié (cf. Figure 5.5). L'approche atteint une micro-F1 de 80,7 % sur ALL_{EN}.
.

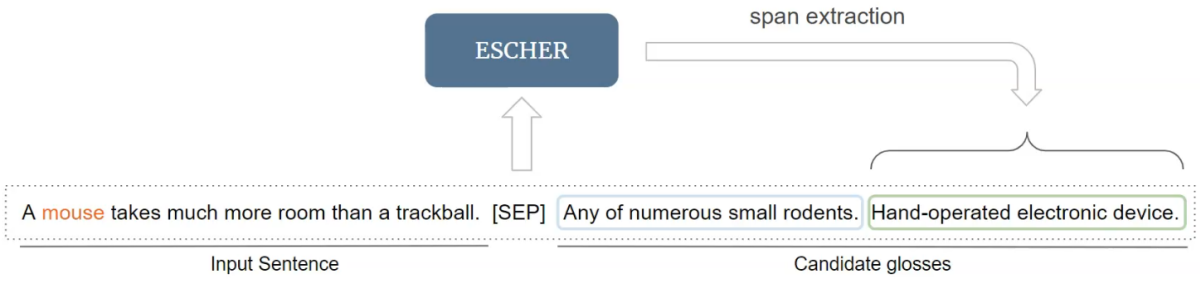


FIGURE 5.5 – Schéma du modèle `ESCHER` pour l'extraction de séquence (diapositive 10 de la présentation (Barba et al., 2021a)).

Enfin, le modèle aujourd'hui le plus performant pour la désambiguïsation lexicale de l'anglais est `ConSeC` (CONtinuous SEnse Comprehension) (Barba et al., 2021b) dont l'idée centrale est de considérer la tâche comme un **processus continu et itératif**. `ConSeC` désambiguïse tous les lexèmes ambigus d'une phrase à l'aide d'un mécanisme de boucle de rétroaction. Au lieu de traiter chaque lexème-cible indépendamment, le modèle classe les lexèmes par ordre croissant de polysémie et les désambiguïse un par un, en se basant à chaque fois sur les sens déjà attribués aux lexèmes précédemment traités.

Pour chaque lexème-cible, le système alimente un modèle de type transformeur (`DEBERTA`) avec la phrase d'entrée, les définitions candidates pour ce lexème et des définitions des lexèmes voisins déjà désambiguïsés, qui sont insérés dans la séquence. Le modèle est entraîné comme un modèle d'extraction de texte, prédisant quelle séquence de texte correspond le mieux à la signification du lexème, reprenant l'idée appliquée dans `ESCHER`. Cette approche a permis d'améliorer les performances de `ESCHER`, atteignant une micro-F1 de 83,2 %.

5.4.3 Approches avec les grands modèles de langue auto-régressifs

Depuis le développement des grands modèles de langue auto-régressifs, plusieurs travaux se sont intéressés à leurs capacités à résoudre une tâche de désambiguïsation lexicale.

Kocoń et al. (2023) ont testé `ChatGPT 3.5` sur diverses tâches dont la tâche de désambiguïsation lexicale et de *word-in-context* (cf. Section 5.7.1). Le modèle est testé sur ALL_{EN} (Raganato et al., 2017) avec des invites (*prompts*) contenant une phrase avec un lexème-cible et la liste de ses sens possibles décrits avec une courte définition. Le modèle doit ensuite générer l'ID du sens choisi. Dans la mesure du possible, les résultats sont comparés aux résultats état-de-l'art sur chaque tâche. `ChatGPT 3.5` produit une macro-F1 de 73,30 %, à comparer au 83,20 % du modèle `ConSeC` (Barba et al., 2021b). Cependant, `ConSeC` est vraisemblablement évalué avec une micro-F1, traditionnellement utilisée dans la tâche de désambiguïsation lexicale, empêchant une réelle comparaison.

Une difficulté dans l'utilisation des grands modèles de langue est leur **forte sensibilité au *prompt*** et à la manière dont les informations sont formulées (Errica et al., 2025; Sahoo et al., 2025). Un large éventail de stratégies de *prompting* doit donc être testé.

Pour la tâche de désambiguisation lexicale, Sumanathilaka et al. (2024) explore le *Chain-of-Thought prompting* (Wei et al., 2022) et l'augmentation de *prompt* avec l'intégration d'informations linguistiques, telles que les étiquettes de partie du discours ou des synonymes des lexèmes à forte ambiguité. Des grands modèles de langue comme `Llama-2-70B`, `Llama-3-70B`, `GPT4-Turbo` ou `Gemma-7B` sont testés sur le jeu de données `FEWS` (Blevins et al., 2021) qui présente davantage de lexèmes ambigus que ALL_{EN}. `LLama-2-70B` produit les meilleures performances avec 83 % d'exactitude contre 79,06 % avec la meilleure *baseline* dans l'article d'origine.

Au-delà de demander aux grands modèles de langue de choisir le sens correct parmi une liste de sens donnée, plusieurs travaux ont exploité les **capacités de génération textuelle de ces modèles** en récupérant une définition générée et la comparant à la définition correcte.

Basile et al. (2024) proposent un *benchmark* sur l'italien avec des données provenant du *framework* `XL-WSD` et comparent les définitions générées à des définitions *gold* avec deux métriques : F1-RougeL et F1-BERTscore. Sur la tâche de génération de définitions, les trois modèles choisis (`LLama3.1-8B-Instruct`, `Mistral-7B-Instruct` et `Gemma2-9B-Instruct`) montrent des scores

RougeL faibles (en moyenne 0,1) et des BERTscores plus élevés (en moyenne 0,68). Sur la configuration classique, le meilleur modèle produit une exactitude de 58,40 %.

Dans leur extension à d'autres langues (anglais, français, espagnol et allemand) et à d'autres grands modèles de langue, les résultats montrent que les grands modèles de langue les plus larges (plus de 70 milliards de paramètres) produisent de meilleures performances, notamment sur l'anglais (Basile et al., 2025).

Entraînés sur une quantité massive de données provenant du *web*, les grands modèles de langue présentent un **problème connu de contamination de données**, c'est-à-dire la présence potentielle de données d'évaluation dans les données d'entraînement (Balloccu et al., 2024).

Ce phénomène étant difficilement vérifiable en raison de l'opacité des modèles concernant leurs données d'entraînement, Meconi et al. (2025) produisent un nouveau jeu de données d'évaluation pour la tâche de désambiguïsation lexicale avec des grands modèles de langues, `WikiPortal WSD`. Les modèles sont également évalués sur les jeux de données ALL_{NEW}, `hardEN` et `42D` (Maru et al., 2022). Leur étude évalue une quinzaine de grands modèles de langue, à la fois *open* et *close-weight* avec 20 *prompts* distincts, afin de tester différentes stratégies de *prompting*. Le tableau 5.2 présente deux exemples de *prompt* : un pour la désambiguïsation sous la forme de tâche de classification et un pour la génération de définition. Ces *prompts* sont ensuite dérivés selon la stratégie de *prompting* adoptée : formulation en questions directes, format d'un dialogue, génération de synonymes etc.

Tâche	**Exemple de *prompt***
Classification	Read the sentence : I heard parts of the building above my head cracking. Choose the correct dictionary definition of the word “building" from the options : 1. A structure that has a roof and walls and stands more or less permanently in one place. 2. The act of constructing something. 3. The commercial activity involved in repairing old structures or constructing new ones. 4. The occupants of a building. Provide as output only the correct dictionary definition.
Génération de définition	Consider the following sentence : they needed rugs to cover the bare floors. Provide a dictionary-style definition which identifies the meaning of “floor" in the above sentence. Do not motivate your answer and do not refer to the sentence context

Table 5.2 – Exemples de *prompts* testés par Meconi et al. (2025).

Les résultats montrent que les modèles `GPT-4o` et `DeepSeek-V3` parviennent à des performances proches des modèles état-de-l'art tels que `ConSeC` et `ESCHER` sur ALL_{NEW} et le `WikiPortal WSD` (micro-F1 autour de 83 %). Toutefois, les performances des grands modèles de langue se démarquent en particulier sur les jeux de données plus complexes comme `hardEN` (micro-F1 de 45,6 % pour `GPT-4o` contre 7,4 % pour `ConSeC`) et `42D` (micro-F1 de 76,8 % pour `GPT-4o` contre 58,9 % pour `ESCHER`), présentant ainsi une meilleure adaptabilité aux instances provenant de différents domaines (`42D`) et à celles généralement mal prédites par les systèmes

supervisés (`hardEN`).

5.5 Métriques traditionnelles

Les métriques traditionnelles pour évaluer la tâche de désambiguïsation sont la **précision**, le **rappel** et la **F1** (F-mesure), généralement utilisées dans les tâches de classification, que nous détaillerons dans le contexte de la tâche de désambiguïsation lexicale.

Classification binaire. Dans une tâche de classification binaire, la précision et le rappel sont définis tels que :

$$Precision = \frac{VP}{VP + FP} \qquad Rappel = \frac{VP}{VP + FN}$$

où VP = Vrai Positif, FP = Faux Positif et FN = Faux Négatif.

La F1 (ou F-mesure) est la moyenne harmonique de la précision et du rappel :

$$F1 = \frac{2}{Rappel^{-1} + Precision^{-1}} = \frac{2 \times VP}{2 \times VP + FP + FN}$$

Classification à classes multiples. La tâche de désambiguïsation lexicale est une classification à classes multiples, les différentes classes correspondant aux sens possibles d'un lexème à désambiguïser.

Pour calculer la F1 dans cette configuration, une première approche est de faire la moyenne des F1 de chaque classe, correspondant à la **macro-F1**.

$$macro\text{-}F1 = \frac{2}{macroRappel^{-1} + macroPrecision^{-1}}$$

avec

$$macroRappel = \frac{1}{\#classes} \sum_{C \in classes} Rappel_C$$

$$macroPrecision = \frac{1}{\#classes} \sum_{C \in classes} Precision_C$$

On adopte l'approche « un contre le reste » (One-vs-Rest) pour revenir à une classification binaire. On reprend l'exemple du lexème « avocat » qui possède les trois sens suivants :

— C_1 : avocat_fruit
— C_2 : avocat_métier
— C_3 : avocat_couleur

Pour la classe C_1 - avocat_fruit, on récupère le nombre de VP, FP et FN, comme décrit dans le tableau 5.3. On peut ainsi calculer la précision, le rappel et la F1 pour chaque classe et faire la moyenne des F1 pour obtenir la macro-F1.

Une autre approche consiste à calculer le nombre total de vrais positifs (VP), faux positifs (FP) et faux négatifs (FN) pour toutes les instances, sans distinction de classe. On calcule alors une **micro-F1**.

prédiction	*gold*	résultat
C_1	C_1	VP
reste	C_1	FN
C_1	reste	FP

TABLE 5.3 – Exemple de classification à classes multiples avec l'approche « un contre le reste », « reste » signifiant n'importe quelle autre classe.

$$\textit{micro-F1} = \frac{2}{\textit{microRappel}^{-1} + \textit{microPrecision}^{-1}}$$

avec

$$\textit{microPrecision} = \frac{VP}{VP + FP} \qquad \textit{microRappel} = \frac{VP}{VP + FN}$$

Contrairement à la macro-F1, une F1 n'est pas calculée pour chaque classe ; on calcule le total de vrais positifs, faux positifs et faux négatifs (cf. Tableau 5.4). Une prédiction du modèle différente du *gold* est considérée comme un cas de faux positif ET de faux négatif. Cette configuration revient à obtenir la même valeur pour la précision, le rappel et la micro-F1.

prédiction	*gold*	résultat
C_i	C_i	VP
C_i	C_j, $j \neq i$	FN / FP

TABLE 5.4 – Exemple d'un calcul d'une micro-F1 dans une classification à classes multiples.

Par ailleurs, une exactitude (*accuracy*) peut également être calculée. Dans le cas où le jeu de données implique un seul sens correct par instance, l'exactitude est égale à la micro-F1.

$$\textit{Exactitude} = \frac{\#\text{ prédictions correctes}}{\#\text{ total de prédictions}}$$

La majorité des travaux sur la désambiguïsation évalue les performances des modèles avec une **micro-F1**. Il est également courant de trouver l'exactitude dans certains travaux (Sinha et al., 2022; Sumanathilaka et al., 2024; Basile et al., 2024).

5.6 Une évaluation biaisée par un cadre avantageux

Durant les 20 dernières années, la manière d'évaluer les systèmes de désambiguïsation lexicale s'est ancrée dans un schéma fixe : calcul des performances avec une micro-F1 sur des *frameworks* de référence, les données ALL_{EN} (Raganato et al., 2017) pour l'anglais, `XL-WSD` (Pasini et al., 2021) pour la configuration multilingue, qui a toutefois été peu explorée.

Si cette configuration permet une comparaison constante entre les modèles (cf. Tableau 5.5) et donne l'impression d'atteindre l'accord inter-annotateurs estimée à 80 % pour une désambiguïsation avec un inventaire de sens à granularité fine (Navigli, 2009), elle ne reflète pas les performances réelles des modèles. En effet, Maru et al. (2022) mettent en avant les problématiques de l'évaluation de la désambiguïsation lexicale, notamment celles liées aux biais des données et à l'utilisation unique de la micro-F1.

Dans cette section, nous nous proposons de détailler ces problématiques en reprenant les analyses et les résultats présentés dans Maru et al. (2022). Contrairement aux auteurs qui proposent des analyses sur les différences de performance entre les modèles, nous nous concentrons davantage sur la variabilité du score de performance selon les données d'évaluation. Notre objectif principal est de montrer qu'une métrique unique ne suffit pas à démontrer qu'un système est réellement performant et qu'il est nécessaire de fournir une analyse qualitative des résultats.

Nous complétons ainsi leur étude avec de nouveaux résultats que nous mettrons en évidence en gris dans les tableaux.

5.6.1 Biais de la composition des données d'évaluation

Le tableau 5.5 présente les performances en micro-F1 des modèles que nous avons détaillés en section 5.4, sur les jeux de données décrits en section 5.3. Les chiffres sont issus des articles d'origine.

	ALL_{EN}	ALL_{NEW}	FEWS	hardEN	42D
MFS *baseline*	64,8	65,2	—	—	—
SyntagRank (Scozzafava et al., 2020)	71,7	—	—	0,0	28,1
$SREF_{KB}$ (Wang and Wang, 2020)	73,5	—	—	—	—
EWISE (Kumar et al., 2019)	71,8	—	—	—	—
LMMS (Loureiro and Jorge, 2019)	75,4	—	—	—	—
GlossBERT (Huang et al., 2019)	77,0	—	—	0,0	41,9
$SREF_{sup}$ (Wang and Wang, 2020)	77,8	—	—	—	—
BEM (Blevins and Zettlemoyer, 2020; Blevins et al., 2021)	79,0	—	79,1	0,0	47,8
EWISER (Bevilacqua and Navigli, 2020)	80,1	79,2	—	0,0	40,8
ESCHER (Barba et al., 2021a)	80,7	81,6	—	0,0	54,1
ConSeC (Barba et al., 2021b)	83,2	83,0	—	7,4	56,6
Llama-2 70b (Sumanathilaka et al., 2024)	—	—	83,0	—	—
GPT-4o (Meconi et al., 2025)	—	82,3	—	45,6	76,8
DeepSeek-V3 (Meconi et al., 2025)	—	81,9	—	39,5	74,6

TABLE 5.5 – Performances (%) en micro-F1 des principaux modèles de désambiguisation lexicale.

La majorité des systèmes ont été évalués sur le jeu de données ALL_{EN} de Raganato et al. (2017), qui a longtemps été le jeu de données de référence pour la désambiguïsation lexicale de l'anglais. On remarque que les performances augmentent jusqu'à dépasser le « plafond de verre » de 80 % de micro-F1 mentionné par Bevilacqua and Navigli (2020).

Bien que les résultats sur ALL_{EN} donnent l'impression d'être proches d'une performance parfaite, on note que les scores sont bien plus faibles sur d'autres jeux de données tels que hardEN et 42D, constitués d'instances plus complexes, témoignant de l'importance d'évaluer les systèmes

sur des données variées.

Le jeu de données d'évaluation ALL_{EN} est constitué de 7 253 instances où des noms communs, verbes, adjectifs et adverbes sont annotés en sens `WordNet` (Raganato et al., 2017). Il contient des instances monosémiques et des instances présentant plusieurs sens corrects (*multi-gold*) en raison de la granularité fine de `WordNet`. En proportion, la composition des instances de ALL_{EN} est de :

— 17,9 % d'instances monosémiques (1 301),

— 77,4 % d'instances polysémiques avec un sens correct (5 612),

— 4,7 % d'instances polysémiques avec plusieurs sens corrects (322 avec deux sens corrects et 18 avec trois sens corrects.).

Les instances monosémiques. La présence d'instances monosémiques dans le jeu de données d'évaluation se révèle être peu cohérente avec le format de la tâche de désambiguïsation lexicale dans la mesure où le système choisit le sens correct parmi les sens possibles du lexème à désambiguïser. De ce fait, le système prédit automatiquement le sens correct pour ces instances monosémiques, tirant ainsi la micro-F1 vers le haut.

Afin d'observer le poids des instances monosémiques dans la micro-F1 finale, nous déduisons la micro-F1 sur les instances polysémiques, la micro-F1 pour les instances monosémiques étant de 100 % :

$$microF1_{poly} = \frac{microF1_{totale} - \%mono}{\%poly}$$

où $\%mono$ et $\%poly$ sont respectivement les pourcentages d'instances monosémiques et d'instances polysémiques.

Le tableau 5.6 présente les performances de cinq systèmes [53] sur $ALL_{ENnoMono}$ correspondant à ALL_{EN} sans les instances monosémiques. On remarque qu'en enlevant les 1 301 instances monosémiques, les performances diminuent de quatre à six points.

Données	**# instances**	`SyntagRank`	`GlossBERT`	`BEM`	`EWISER`	`ESCHER`
ALL_{EN}	7 253	71,7	76,9	79,0	80,1	80,7
$ALL_{ENnoMono}$	5 592	65,5	71,8	74,4	75,8	76,5

TABLE 5.6 – Performances en micro-F1 sur ALL_{EN} et $ALL_{ENnoMono}$. Nous avons ajouté les résultats de la ligne en gris.

Les instances *multi-gold*. Maru et al. (2022) analysent les instances de ALL_{EN} que les systèmes les plus performants ne parviennent pas à désambiguïser. Avec l'aide d'un linguiste, ils identifient les raisons qui expliquent les difficultés des systèmes à désambiguïser ces instances et les corrections nécessaires à appliquer. Les annotations montrent ainsi :

53. Les cinq systèmes considérés constituent l'intersection des systèmes décrits en section 5.4.2 et ceux présents dans Maru et al. (2022).

— 9,4 % d'instances où un sens correct devrait être ajouté à l'ensemble des sens corrects présents dans la ressource,

— 8 % d'instances présente un ou plusieurs sens correct(s) erronés,

— 6,8 % d'instances présente un sens correct erroné mais il n'existe pas de sens approprié dans `WordNet 3.0`,

— 2,9 % d'instances présente une erreur dans le lemme ou la tokenisation,

— 0,3 % d'instances avec une erreur dans l'étiquette de partie du discours,

— 72,6 % d'instances qui ne nécessitent pas de corrections.

Après une mise à jour et une correction des instances, le jeu de données ALL_{NEW} est créé, contenant 4 917 instances toutes polysémiques. Le tableau 5.7 présente les résultats sur ce nouveau jeu de données et les instances correspondantes dans ALL_{EN} (avant correction).

Données	**#instances**	***(#multi-gold)***	`SyntagRank`	`GlossBERT`	`BEM`	`EWISER`	`ESCHER`
$ALL_{EN}*$	4 917	(283)	66,9	74,4	76,2	76,0	78,3
ALL_{NEW}	4 917	(599)	68,5	77,4	79,5	79,2	81,6

TABLE 5.7 – Performances (%) en micro-F1 sur $ALL_{EN}*$ et ALL_{NEW}. $ALL_{EN}*$ correspond au sous-ensemble de ALL_{EN} dont les instances sont également dans ALL_{NEW} (Tableau 5 de Maru et al. (2022)). *#multi-gold* correspond au nombre d'instances présentant plusieurs sens corrects.

On peut noter que les systèmes présentent de meilleures performances sur ALL_{NEW}, ce qui suggère que les corrections appliquées vont en faveur des systèmes, notamment dans les cas où le système prédisait le sens correct mais que celui-ci était en réalité erroné.

Par ailleurs, nous calculons le nombre d'instances *multi-gold* dans $ALL_{EN}*$ et ALL_{NEW}. Ce nombre passe de 283 dans $ALL_{EN}*$ à 599 dans ALL_{NEW}, soit plus du double.

En observant le nombre d'instances *multi-gold* en fonction du nombre exact de sens corrects dans $ALL_{EN}*$ et ALL_{NEW} (cf. Tableau 5.8), on remarque que des instances peuvent présenter jusqu'à quatre sens corrects dans ALL_{NEW}, contre maximum trois dans $ALL_{EN}*$. Le nombre d'instances *2-gold* est presque doublé et le nombre d'instances *3-gold* est cinq fois plus élevé après la mise à jour des données.

L'augmentation d'instances *multi-gold* explique également les meilleures performances observées sur ALL_{NEW} dans la mesure où l'évaluation classique considère le cas où un système prédit un sens correct parmi l'ensemble des sens corrects comme un vrai positif. De ce fait, attribuer plusieurs sens corrects à un lexème permet d'alléger la tâche de désambiguïsation.

Données	# inst *multi-gold*	# inst *2-gold*	# inst *3-gold*	# inst *4-gold*
$ALL_{EN}*$	283	266	17	0
ALL_{NEW}	599	465	86	48

TABLE 5.8 – Statistiques sur les instances *multi-gold* dans $ALL_{EN}*$ et ALL_{NEW}.

5.6.2 Biais de la métrique

La majorité des travaux utilisent une micro-F1 pour évaluer les systèmes de désambiguïsation lexicale. Cependant, la micro-F1 et la macro-F1 ne mesurent pas exactement le même phénomène. Avec une macro-F1, les classes ont un poids identique dans la moyenne, indépendamment du nombre d'instances par classe, contrairement à la micro-F1.

Pour illustrer cette différence, un cas artificiel extrême de deux classes non équilibrées « avocat_fruit » et « avocat_métier » est présenté dans le tableau 5.9. On créé un scénario où la classe la plus fréquente « avocat_fruit » est systématiquement prédite correctement et une classe peu fréquente « avocat_métier » est incorrectement prédite.

Classe	**# instances**	**VP**	**FP**	**FN**
avocat_fruit	99	99	1	0
avocat_métier	1	0	0	1

TABLE 5.9 – Exemple de résultats de classification avec des classes non équilibrées.

Pour la macro-F1, on calcule les F1 pour chaque classe :

$$P_{\text{fruit}} = \frac{99}{99+1} = 0,99 \qquad R_{\text{fruit}} = \frac{99}{99+0} = 1,0$$

$$F1_{\text{fruit}} = \frac{2 \times VP}{2 \times VP + FP + FN} = \frac{2 \times 99}{2 \times 99 + 1 + 0} = 0,995$$

$$F1_{\text{metier}} = 0$$

$$\textit{macro-F1} = \frac{F1_{\text{fruit}} + F1_{\text{metier}}}{2} = \frac{0,995 + 0}{2} = 0,4975$$

Pour la micro-F1, on compte le nombre de VP, FP et FN sur toutes les instances pour obtenir la *microPrecision*, le *microRappel* et la *micro-F1* :

$$VP = 99 \qquad FP = 1 \qquad FN = 1$$

$$\textit{microPrecision} = \frac{VP}{VP + FP} = \frac{99}{100} = 0,99 \qquad \textit{microRappel} = \frac{VP}{VP + FN} = \frac{99}{100} = 0,99$$

$$\textit{micro-F1} = \frac{2 \times \textit{microPrecision} \times \textit{microRappel}}{\textit{microPrecision} + \textit{microRappel}} = \frac{2 \times 0,99 \times 0,99}{0,99 + 0,99} = 0,99$$

Dans ce cas très déséquilibré où une classe très fréquente est très bien prédite par le système, calculer une micro-F1 donne un résultat très élevé à 99,0 %. Cependant, avec une macro-F1 qui donne le même poids à toutes les classes indépendamment du nombre d'instances, le système présente une performance bien plus faible, à 49,75 %.

Le tableau 5.10 présente les performances de cinq systèmes de désambiguïsation lexicale avec une micro-F1 (m-F1) et une macro-F1 (M-F1) sur ALL_{EN}, $ALL_{ENnoMono}$ et ALL_{NEW}.

Nous déduisons la macro-F1 sur $ALL_{ENnoMono}$ avec la formule suivante :

$$macroF1_{poly} = \frac{macroF1_{totale} - \%mono}{\%poly}$$

où $\%mono$ et $\%poly$ sont respectivement les pourcentages d'instances monosémiques et d'instances polysémiques.

Dans un premier temps, on observe que les performances en macro-F1 diminuent par rapport à celles en micro-F1. L'écart entre la micro-F1 et la macro-F1, représentée par le delta Δ, est variable selon les modèles et les jeux de données.

Sur ALL_{EN}, l'écart est entre 4,3 et 7,6 points selon les modèles. Plus le Δ est élevé, plus il y a de sens mal prédits par le système, chaque sens ayant le même poids indépendamment de sa fréquence d'apparition. Les écarts sont logiquement plus élevés sur $ALL_{ENnoMono}$ (entre 7,1 et 12,1) car les instances monosémiques que les systèmes prédisent systématiquement correctement, sont retirées.

Données	**#inst**	SyntagRank			GlossBERT			BEM			EWISER			ESCHER		
		M-F1	m-F1	Δ	M-F1	m-F1	Δ	M-F1	m-F1	Δ	M-F1	m-F1	Δ	M-F1	m-F1	Δ
ALL_{EN}	7 253	64,1	71,7	7,6	71,3	76,9	5,6	73,9	79,0	5,1	75,8	80,1	4,3	73,3	78,3	5,0
$ALL_{ENnoMono}$	5 952	53,4	65,5	12,1	62,7	71,8	9,1	66,1	74,4	8,3	65,3	73,6	8,3	69,4	76,5	7,1
ALL_{NEW}	4 917	61,4	68,5	7,1	73,2	77,4	4,2	75,6	79,5	3,9	75,6	79,2	3,6	78,7	81,6	2,9

TABLE 5.10 – Performances (%) en macro-F1 (M-F1) et micro-F1 (m-F1), avec deltas (Δ = m-F1 - M-F1), sur ALL_{EN}, $ALL_{ENnoMono}$ et ALL_{NEW} (Tableau 1 de Maru et al. (2022)).

S'appuyer sur plusieurs métriques est d'autant plus important dans le contexte de la tâche de désambiguïsation lexicale où les systèmes peuvent être influencés par la distribution des sens dans les jeux de données.

5.6.3 Biais du sens le plus fréquent et des données d'entraînement

Un des défis majeurs dans la tâche de désambiguïsation lexicale est l'influence de la fréquence d'apparition des sens dans les données. Dans un texte naturel, la distribution des sens suit généralement une loi de Zipf : un faible nombre de sens représente la majorité des occurrences tandis que les autres sens apparaissent très rarement.

En pratique, ce phénomène entraîne un biais. Les systèmes de désambiguïsation lexicale ont tendance à prédire le sens le plus fréquent de chaque lexème (Postma et al., 2016). Ce biais peut s'expliquer par la forte représentation des sens les plus fréquents dans les données d'entraînement. Inversement, les sens rares, peu connus lors de la phase d'entraînement, sont souvent mal prédits. Une *baseline* classique consiste alors à attribuer à un lexème-cible son sens le plus fréquent : *Most Frequent Sense baseline* (MFS) (Navigli, 2009).

Sens le plus fréquent. Maru et al. (2022) conduisent une analyse quantitative de la tâche de désambiguïsation lexicale, en mettant en valeur la performance des modèles sur les sens les plus rares. Dans un premier temps, ils isolent les 541 instances de ALL_{EN} que les systèmes les

plus performants ne parviennent pas à désambiguïser, i.e. la performance est de 0 % sur ce jeu de données. Ces instances constituent le jeu de données ALL_{HC} (HC pour *Hard Core*).

Le tableau 5.11 présente le pourcentage de fois que les systèmes prédisent le sens le plus fréquent sur ALL_{EN} et ALL_{HC}.

Données	**#inst (#mono)**	SyntagRank	GlossBERT	BEM	EWISER	ESCHER	**gold**
ALL_{EN}	7 253 (1 301)	81,1	74,8	72,6	72,7	71,2	65,2
ALL_{HC}	541 (0)	80,2	70,6	71,0	67,8	68,6	2,0

TABLE 5.11 – Pourcentage de fois que les systèmes prédisent le sens le plus fréquent (Tableau 1 de Maru et al. (2022)).

On remarque que les systèmes présentent effectivement un biais vers le sens le plus fréquent qui est prédit avec un pourcentage élevé, au moins 71 % contre un pourcentage réel de 65,2 % sur ALL_{EN}. Les chiffres sont toutefois biaisés par les 1 301 instances monosémiques pour lequel le modèle prédit forcément le sens correct.

Le phénomène est nettement plus marqué sur ALL_{HC} avec un pourcentage d'au moins 62 % tandis que le pourcentage réel est de 2 %.

Données d'entraînement. Une autre source de biais est la présence (ou l'absence) des sens dans les données d'entraînement. En effet, prédire un sens qui n'apparaît jamais dans la phase d'entraînement requiert des capacités *zero-shot* aux modèles. Un autre risque est que le modèle favorise d'autres sens du lexème qui ont été vus durant l'entraînement.

Pour mesurer ce biais, Maru et al. (2022) calculent le pourcentage de fois où les systèmes prédisent un sens qui apparaît au moins une fois dans les données d'entraînement (cf. Tableau 5.12).

Sur ALL_{EN}, on note que les systèmes prédisent un sens présent dans les données d'entraînement au moins 86,3 % de fois contre un pourcentage réel de 84,3 %, ce qui montre un léger biais.

Sur ALL_{HC}, le biais est davantage marqué avec au moins 96,5 % de fois contre un pourcentage réel de 67,1 %.

Données	**#inst (#mono)**	**SyntagRank**	GlossBERT	BEM	EWISER	ESCHER	**gold**
ALL_{EN}	7 253 (1 301)	88,8	88,6	87,4	88,8	86,3	84,3
ALL_{HC}	541 (0)	98,3	97,2	96,7	98,0	96,5	67,1

TABLE 5.12 – Pourcentage de fois que les systèmes prédisent un sens qui apparaît au moins une fois dans l'échantillon d'entraînement (Tableau 1 de Maru et al. (2022)).

5.6.4 Des performances dépendantes des données

Dans les sections 5.6.1, 5.6.2 et 5.6.3, nous avons discuté de quatre sources de biais influençant le score de performance des systèmes de désambiguïsation lexicale :

— **biais-eval** : la composition des données d'évaluation,

— **biais-metrique** : le choix de la métrique,

— **biais-mfs** : le biais du sens le plus fréquent,
— **biais-train** : la distribution des sens dans les données d'entraînement.

Afin d'observer la variation des performances des systèmes selon ces quatre sources de biais, nous présentons dans le tableau 5.13, une vue d'ensemble des résultats en micro-F1 et macro-F1 (**biais-metrique**) sur différents jeux de données :

— ALL_{EN} : le jeu de données d'évaluation original de Raganato et al. (2017).
— $ALL_{ENnoMono}$: ALL_{EN} sans les instances monosémiques (**biais-eval**).

— ALL_{ENnoSC} : ALL_{EN} sans les instances dont le sens correct est dans `SemCor` (**biais-train**).
— $ALL_{ENnoSCnoMono}$: ALL_{EN} sans les instances monosémiques ni les instances dont le sens correct est dans `SemCor` (**biais-eval** et **biais-train**).

— ALL_{no1st} : ALL_{EN} sans les instances dont le sens correct est le premier sens dans `WordNet`, i.e. le sens le plus fréquent (**biais-mfs**).

— `42D` : jeu de données d'instances polysémiques dont le sens correct n'est ni le sens le plus fréquent, ni un sens présent dans `SemCor` (**biais-eval**, **biais-train** et **biais-mfs**). Ce jeu de données a été explicitement créé pour évaluer les systèmes dans une situation davantage impartiale (Maru et al., 2022).

Données	**#inst (#mono)**	**SyntagRank**			**GlossBERT**			**BEM**			**EWISER**			**ESCHER**		
		M-F1	m-F1	Δ	M-F1	m-F1	Δ	M-F1	m-F1	Δ	M-F1	m-F1	Δ	M-F1	m-F1	Δ
ALL_{EN}	7 253 (1 301)	64,1	71,7	7,6	71,3	76,9	5,6	73,9	79,0	5,1	75,8	80,1	4,3	76,4	80,7	4,3
$ALL_{ENnoMono}$	5 952 (0)	53,4	65,5	12,1	62,7	71,8	9,1	66,1	74,4	8,3	65,3	73,6	8,3	69,4	76,5	7,1
ALL_{ENnoSC}	1 138 (448)	55,1	61,0	5,9	57,4	62,2	4,8	63,7	67,1	3,4	58,6	64,0	5,4	71,0	75,0	4,0
$ALL_{ENnoSCnoMono}$	690 (0)	—	35,7	—	—	37,6	—	—	45,7	—	—	40,6	—	—	58,8	—
$ALL_{ENno1st}$	2 525 (0)	26,9	29,5	2,6	42,4	45,0	2,6	47,8	50,5	2,7	46,8	49,0	2,2	54,2	55,2	1,0
42D	370 (0)	32,8	28,1	-4,7	45,7	41,9	-3,8	53,2	47,8	-5,4	43,9	40,8	-3,1	58,9	54,1	-4,8

TABLE 5.13 – Performances (%) en macro-F1 (M-F1) et micro-F1 (m-F1), avec deltas (Δ = m-F1 - M-F1) des systèmes sur ALL_{EN}, $ALL_{ENnoMono}$, ALL_{ENnoSC}, $ALL_{ENnoSCnoMono}$, $ALL_{ENno1st}$ et 42D (Tableau 4 de Maru et al. (2022)).

biais-metrique. Sur le jeu de données ALL_{EN} et ses sous-ensembles, les performances en macro-F1 sont plus faibles qu'avec une micro-F1, ce qui indique que les systèmes sont moins performants sur les classes peu fréquente. Cependant, sur `42D`, les performances en macro-F1 sont plus élevées avec par exemple 58,9 % en macro-F1 pour `ESCHER` contre 54,1 % avec une micro-F1.

Ce résultat s'explique par les différences dans la distribution des sens dans les données d'évaluation entre ALL_{EN} et `42D` (cf. Figure 5.6 et tableau 5.14).

Dans ALL_{EN}, si 97,7 % des sens apparaissent au maximum dans 10 instances, il existe des sens apparaissant dans plus d'une vingtaine d'instances, au maximum 61 instances. Le nombre d'instances par sens est plus variable, avec des sens bien plus fréquents que d'autres, ce qui influence la micro-F1.

À l'inverse, dans 42D, 78,9 % des sens apparaissent dans une instance et un sens apparaît au maximum dans cinq instances. Dans la mesure où il existe en moyenne une instance par sens, un sens (ou une classe) n'est pas avantagé par rapport à sa fréquence d'apparition, ce qui peut expliquer pourquoi les performances en macro-F1 sont plus élevées que celles en micro-F1.

Données	**#instances**	**# sens**	**# moyen d'instances par sens**	**# max d'instances par sens**
ALL_{EN}	7 253	3 363	2,26	61
42D	370	338	1,19	5

TABLE 5.14 – Statistiques sur la distribution des sens dans ALL_{EN} et 42D.

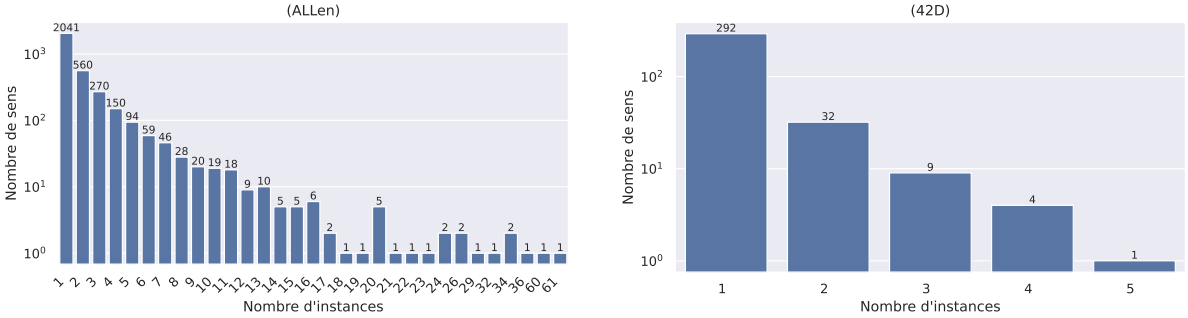


FIGURE 5.6 – Distribution du nombre d'instances par sens dans ALL_{EN} et 42D.

biais-train et biais-eval. Le jeu de données ALL_{ENnoSC} qui contient uniquement les instances de ALL_{EN} dont le sens correct n'est pas dans SemCor, constitue seulement 15 % des instances de ALL_{EN}.

Sur les 1 138 instances de ALL_{ENnoSC}, les systèmes présentent des scores plus faibles que sur ALL_{EN}, le système le plus performant ESCHER, passant d'une micro-F1 de 80,7 % à 75,0 %. Si la diminution des performances a l'air relativement faible, il est toutefois biaisé par les instances monosémiques dans ALL_{ENnoSC} qui constitue environ 39 % des instances. En retirant ces instances, $ALL_{ENnoSCnoMono}$, les performances diminuent fortement avec une micro-F1 de 58,8 % pour ESCHER, soit 21,9 points de différence.

biais-mfs et biais-metrique. Le jeu de données $ALL_{ENno1st}$ contient les instances de ALL_{EN} dont le sens correct n'est pas le premier sens WordNet, les sens étant classés selon leur fréquence d'apparition dans SemCor (le premier sens correspond au sens le plus fréquent). Ces instances constituent environ 35 % de ALL_{EN}.

Les systèmes présentent des performances plus faibles sur ces instances, avec le modèle le plus performant, ESCHER, passant d'une micro-F1 de 80,7 % sur ALL_{EN} à 55,2 % sur $ALL_{ENno1st}$.

On remarque également que l'écart entre la macro-F1 et la micro-F1 est moins marqué sur ce jeu de données, en moyenne 2,22 points. Les instances dont le *gold* est le sens le plus fréquent étant retirés, on peut supposer que les instances dans $ALL_{ENno1st}$ présentent des sens plus rares, ce qui impacte moins la macro-F1.

5.7 Conclusion

La désambiguïsation lexicale est une tâche sémantique historique en TAL. Si elle paraît intuitive pour des humains, elle nécessite la prise en compte d'un lexème selon un contexte, difficile à représenter pour un système automatique.

Depuis les débuts du TAL, la tâche a fait l'objet de nombreux travaux notamment grâce aux campagnes d'évaluation Senseval et SemEval, qui ont permis le développement de données et de méthodes. De nombreux systèmes ont ainsi été créés, dans un premier temps basés sur des règles et sur des connaissances, puis sur des techniques d'apprentissage machine, qui exploite des corpus de textes annotés en sens. Les capacités des modèles de langue ont également été explorées récemment sur la tâche. L'évaluation est basée sur des jeux de données et *frameworks* de référence, permettant une comparaison constante des performances des systèmes.

Toutefois, la tâche de désambiguïsation lexicale présente plusieurs limites, tant sur le plan de son format que de son évaluation.

5.7.1 Les limites de la tâche de classification

La configuration traditionnelle de la tâche de désambiguïsation lexicale est de la considérer comme une tâche de classification. Un inventaire de sens décrit les sens possibles d'un lexème et le système automatique doit attribuer le sens correct d'un lexème selon un contexte donné. Si ce format est pratique d'un point de vue computationnel, il présente plusieurs limites :

— un inventaire de sens statique n'ouvre pas la porte à de nouveaux usages des lexèmes,

— une distribution des lexèmes et des sens suivant la loi de Zipf entraîne un fort déséquilibre des classes,

— la difficulté d'obtenir des données annotées en sens par des experts,

— la difficulté d'accéder à des descriptions des sens telles que des définitions, très utilisées dans les modèles supervisés,

— l'incapacité des modèles à généraliser sur des sens rares ou inconnus.

Dans la section 5.4, nous avons décrit les différents systèmes qui se sont succédé au fil des années, les auteurs cherchant à surmonter ces limites et à maximiser les performances.

Les techniques ont alors exploité différentes sources de connaissances telles les dictionnaires (Lesk, 1986; Huang et al., 2019; Kumar et al., 2019), `WordNet` et sa structure (Agirre and Soroa, 2009; Scozzafava et al., 2020; Bevilacqua and Navigli, 2020; Wang and Wang, 2020) et enfin les plongements contextuels provenant de modèles de langue `BERT` (Loureiro and Jorge, 2019; Huang et al., 2019; Blevins and Zettlemoyer, 2020).

La tâche a également été reformulée en différentes tâches telles qu'une tâche d'extraction de séquences (Barba et al., 2021a,b) ou une tâche de génération de définition (Bevilacqua et al., 2020; Basile et al., 2025; Meconi et al., 2025)).

Au-delà de la désambiguïsation lexicale, des variations de la tâche ont été proposées comme la tâche **Word-in-Context (WiC)** (Pilehvar and Camacho-Collados, 2019) qui évalue la compréhension sémantique en demandant aux modèles de déterminer si un lexème est utilisé avec le même sens dans deux contextes différents. La tâche simplifie la construction de jeux de données et prend en charge des *benchmarks* multilingues (Bevilacqua et al., 2021).

Une autre tâche qui ne nécessite pas d'inventaire de sens est la tâche de **substitution lexicale** (McCarthy and Navigli, 2007) dans laquelle les modèles éliminent l'ambiguïté d'un lexème en proposant des substituts qui préservent le sens. Cette tâche s'adapte à l'évolution du vocabulaire et a des applications pratiques telles que l'augmentation des données. Cependant, elle souffre de problèmes de circularité et de cas où il n'existe pas de substituts appropriés.

5.7.2 Un cadre d'évaluation « trop » ancré

Si le *framework* de Raganato et al. (2017) est majoritairement utilisé pour l'évaluation et la comparaison des systèmes, il ne reflète toutefois pas les réelles capacités des modèles et tend à engager les travaux vers une course à la performance. En effet, si le « plafond de verre » de 80 % de performance semble atteint (Bevilacqua and Navigli, 2020) et donne l'impression que les systèmes produisent des résultats proches de l'accord inter-annotateurs humain (Navigli, 2009), l'évaluation des systèmes sur d'autres jeux de données met en évidence que ces chiffres sont largement biaisés (Maru et al., 2022).

En ce sens, nous avons consacré la section 5.6 à l'évaluation des systèmes, mettant en lumière la variabilité des scores de performances selon les données. Plus particulièrement, nous avons illustré l'impact de quatre sources de biais :

— **le biais des données d'évaluation** : la présence d'instances monosémiques et d'instances *multi-gold* augmentent les performances des systèmes.

— **le biais de la métrique** : une micro-F1 tend à « booster » les performances lorsque les sens les plus fréquents sont bien prédits. Dans ce cas de figure, les scores sont généralement bien plus faibles avec une macro-F1.

— **le biais du sens le plus fréquent** : une problématique connue en désambiguïsation lexicale est le fait que les systèmes tendent à prédire le sens le plus fréquent (Postma et al., 2016).

— **le biais des données d'entraînement** : les systèmes prédisent davantage des sens vus durant l'entraînement, leurs capacités à prédire en *zero-shot* pouvant être limitées.

En section 5.6, nous montrons qu'en retirant ces biais, les performances des modèles diminuent drastiquement, le modèle le plus performant passant d'une micro-F1 de 80,7 % sur le jeu de données de référence pendant de nombreuses années, ALL_{EN} contre 54,1 % sur un jeu de données construit intentionnellement dans le but d'être impartial. Cette section vise à montrer qu'un score et un jeu de données d'évaluation uniques sont insuffisants pour évaluer les performances d'un modèle et à prioriser davantage les analyses qualitatives des résultats.

Chapitre 6

Intégration de plongements de graphe et évaluation de la désambiguïsation lexicale d'exemples lexicographiques

Sommaire

6.1 Enjeux et contributions 124
6.2 BEL-RL-fr, un corpus atypique pour la désambiguïsation lexicale 125
6.2.1 Exemples lexicographiques annotés avec le RL-fr 125
6.2.2 Pré-traitement des données 126
6.3 Systèmes de désambiguïsation lexicale 128
6.3.1 *Baselines* 128
6.3.2 Modèle EWISER et intégration de plongements de graphe 131
6.3.3 Grands modèles de langue 133
6.4 Évaluer au-delà de la classification 136
6.4.1 Métriques 136
6.4.2 Évaluation restreinte *vs* Évaluation globale 136
6.5 QR 1 : performances des systèmes en évaluation restreinte . . . 138
6.5.1 Meilleures performances avec un modèle neuronal supervisé 139
6.5.2 Des données d'évaluation qui influencent les métriques 140
6.5.3 Conclusion de la QR 1 141
6.6 QR 2 : influence de l'intégration de connaissances 141
6.6.1 Comparaison des prédictions des modèles 141
6.6.2 Une baisse des scores en évaluation globale 144
6.6.3 Une distribution de probabilités déviée par l'intégration de connaissances 147
6.6.4 Des performances variables selon le niveau de polysémie 149
6.6.5 Conclusion de la QR 2 152
6.7 QR 3 : prédiction des sens non vus à l'entraînement 152
6.7.1 Corrélation entre la distribution des sens à l'entraînement et la performance des modèles 153
6.7.2 Amélioration des performances des sens non vus avec EWISER+A . . 153
6.7.3 Conclusion de la QR 3 154
6.8 QR 4 : analyse des performances des grands modèles de langue 155
6.9 Discussion et conclusion 157

6.1 Enjeux et contributions

Dans ce chapitre, nous présentons des expériences de désambiguïsation lexicale sur des données peu utilisées pour la tâche, le `BEL-RL-fr`, un corpus d'exemples lexicographiques en français annotés avec des sens du réseau lexico-sémantique `RL-fr`. Ces expériences s'inscrivent dans la continuité des travaux de Sinha et al. (2022) qui ont précédemment étudié l'utilisation d'un modèle de désambiguïsation lexicale permettant d'intégrer des informations inhérentes à un réseau lexico-sémantique (Bevilacqua and Navigli, 2020).

Dans ces expériences, nous cherchons dans un premier temps à évaluer différents modèles de désambiguïsation lexicale sur les noms communs et les verbes du `BEL-RL-fr`, notamment un modèle enrichi par les plongements de graphe du `RL-fr` générés dans le chapitre 4. Nous testons également quatre grands modèles de langue en *zero-shot* avec différentes stratégies de *prompting*.

Dans un second temps, au-delà de comparer les performances des systèmes de désambiguïsation lexicale, notre recherche insiste sur le cadre d'évaluation et l'analyse qualitative de différents résultats, que nous aborderons à travers quatre questions de recherches (QR) :

1. Quels systèmes produisent les meilleures performances sur le corpus `BEL-RL-fr` dans un cadre d'évaluation classique ?
2. Quelle est l'influence de l'intégration de connaissances relatives au graphe à travers la matrice d'adjacence dans le modèle `EWISER` ?
3. L'intégration de connaissances dans `EWISER` améliore-t-elle les performances sur des sens non vus lors de l'entraînement ?
4. Quelle est l'influence des stratégies de *prompting* sur les performances des grands modèles de langue ?

Les expériences présentées dans ce chapitre reprennent une partie des travaux de Sinha et al. (2022) que nous adaptons et auxquels nous ajoutons les contributions suivantes :

— l'ajout d'une *baseline* basée sur le voisinage lexical du `RL-fr`,

— l'utilisation de différentes configurations du modèle `EWISER`, dont l'intégration de plongements de graphe.

— l'utilisation de quatre grands modèles de langue,

— la mise en place d'un nouveau cadre d'évaluation, appelé *évaluation globale*,

— une analyse détaillée des résultats à travers l'étude des distributions de probabilités des modèles et des erreurs.

Le chapitre s'articule autour de neuf sections. La section 6.2 présente le corpus du `BEL-RL-fr` ainsi que les pré-traitements appliqués pour la tâche. La section 6.3 détaille les différents systèmes de désambiguïsation. Nous présentons ensuite le cadre et les métriques d'utilisation en section 6.4. Nous abordons ensuite les quatre questions de recherche dans les sections suivantes : la section 6.5 présente la comparaison des performances des systèmes dans un schéma d'évaluation classique, la section 6.6 se focalise sur les modèles `EWISER` avec et sans intégration de connaissances, la section 6.7 présente les résultats sur les sens non vus à l'entraînement et la section 6.8 se concentre sur les performances des grands modèles de langue. Enfin, nous terminons avec la section 6.9 de discussion et de conclusion.

6.2 BEL-RL-fr, un corpus atypique pour la désambiguïsation lexicale

Dans cette section, nous présentons le BEL-RL-fr, un corpus d'exemples lexicographiques annotés avec le réseau lexico-sémantique RL-fr. Initialement conçu pour illustrer des sens du RL-fr, le corpus est atypique dans sa distribution des sens qui ne reflète pas celle d'un texte naturel. Toutefois, les annotations en sens nous permettent d'exploiter le corpus pour une tâche de désambiguïsation sur des données en français.

Nous décrivons dans un premier temps le corpus et l'inventaire de sens, puis nous détaillons les pré-traitements appliqués pour la tache de désambiguïsation lexicale.

6.2.1 Exemples lexicographiques annotés avec le RL-fr

Le BEL-RL-fr (Lux-Pogodalla, 2014) est un corpus d'exemples lexicographiques issus de Frantext [54] et d'articles de l'Est Républicain dont certains lexèmes sont annotés avec des unités lexicales du RL-fr [55]. Des mots de différentes catégories grammaticales sont annotés, telles que des noms communs, des verbes, des adjectifs, des adverbes ou des prépositions. Outre les annotations en unités lexicales du RL-fr, les instances incluent des annotations de caractéristiques grammaticales comme l'étiquette de partie du discours ou des marques d'usage (emploi rare, vieilli, argot, Québec...).

La figure 6.1 présente un exemple d'annotation d'une phrase dans le BEL-RL-fr.

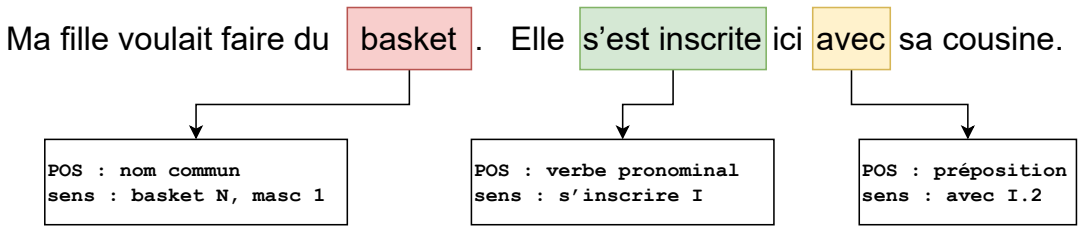


FIGURE 6.1 – Annotation d'un exemple lexicographique dans le BEL-RL-fr v2.0 (exemple 129).

L'inventaire de sens utilisé pour annoter le BEL-RL-fr est le RL-fr, un réseau lexico-sémantique du français dont les nœuds décrivent des unités lexicales, et les arêtes, des relations lexico-sémantiques ou combinatoires (cf. Section 1.3.3 du chapitre 1).

La structure du RL-fr est décomposée en plusieurs concepts utilisant une terminologie précise. Les unités lexicales correspondent aux sens d'un **vocable**, un vocable étant une entrée classique dans un dictionnaire. Par exemple, dans le Wiktionnaire, il existe quatre vocables pour le lexème avocat [56] :

— avocat, nom commun 1 (métier),

— avocat, nom commun 2 (fruit),

— avocat, nom commun 3 (couleur),

54. https://www.frantext.fr/
55. Dans cette thèse, nous avons utilisé les versions 2.0 du BEL-RL-fr RL-fr et 2.1 du RL-fr.
56. https://fr.wiktionary.org/wiki/avocat.

— avocat, adjectif (couleur).

Dans le `RL-fr`, on en compte seulement deux, les noms communs avocat[1] (métier) et avocat[2] (fruit). Chaque vocable peut admettre plusieurs sens, aussi appelés **acceptions** dans le `RL-fr` (cf. Tableau 6.1). Dans ce chapitre, nous utiliserons le terme de **sens** par souci de cohérence avec l'état de l'art que nous avons proposé dans le chapitre 5.

nom lexicographique du vocable (lexème)	avocat	
vocable (entrée)	avocat[1] (métier)	avocat[2] (fruit)
unité lexicale (sens ou acception)	• avocat^1_I : individu qui pratique le métier • avocat$^1_{II}$: individu qui a un certain comportement	avocat[2]

Table 6.1 – Représentation de la polysémie du lexème `avocat` dans le `RL-fr`.

Le `BEL-RL-fr` a été initialement construit pour illustrer les contextes d'apparition des sens du `RL-fr`. Les exemples ayant été choisis manuellement par des lexicographes, la fréquence d'apparition des sens n'est pas représentative de la distribution réelle dans un texte naturel.

S'il n'a pas été originellement créé pour la désambiguïsation lexicale, le corpus comprend des données exploitable pour la tâche dans la mesure où des mots polysémiques sont désambiguïsés avec des sens du `RL-fr` (Lux-Pogodalla, 2014).

Dans ce chapitre, nous présentons les expériences de désambiguïsation lexicale que nous avons menées sur les données du `BEL-RL-fr`. Nous nous concentrons uniquement sur les noms communs et les verbes, qui présentent le plus d'annotations en sens. Le tableau 6.2 présente les statistiques sur les annotations en sens pour ces catégories. On note que 71 % des sens apparaissent une seule fois dans les exemples, avec une occurrence moyenne de 2,14 pour les noms et 1,90 pour les verbes.

	Noms	**Verbes**
Nombre d'exemples	20 783	8 631
Nombre d'instances annotées	28 587	9 950
Nombre de sens utilisés	13 359	5 237
Nombre de sens apparaissant une fois	6 463 (71 %)	3 715 (71 %)
Nombre de sens apparaissant deux fois	1 556 (12 %)	710 (14%)
Nombre de sens apparaissant trois fois	622 (5 %)	285 (5 %)
Nombre moyen d'occurrences d'un sens	2,14	1,90

Table 6.2 – Statistiques sur les exemples lexicographiques du `BEL-RL-fr` (v2).

6.2.2 Pré-traitement des données

Nous avons appliqué les mêmes pré-traitements et stratégie de division en échantillons d'entraînement, de validation et de test que Sinha et al. (2022).

Dans un premier temps, les données sont filtrées pour ne garder que les exemples contenant des verbes ou des noms à désambiguïser. Par ailleurs, tous les exemples dans lesquels un verbe ou

un nom fait partie d'une expression polylexicale, comme des idiomes, ont été retirés, ces exemples ne reflétant pas l'utilisation standard des mots qu'ils contiennent.

Dans le chapitre 4, nous avons généré des plongements de sens sur RLF27k, un sous-ensemble de RL-fr. Afin d'intégrer ces plongements de graphe dans le modèle EWISER (cf. Section 6.3.2), nous retirons les sens qui ne font pas partie de RLF27k, ainsi que les exemples correspondants. Avec ce filtre, 555 instances sont retirées pour les noms et 200 pour les verbes. Au total, 18 965 exemples lexicographiques sont conservés pour les noms communs et 7 231 pour les verbes. Il n'existe pas d'instances annotées avec plusieurs sens (pas d'instances *multi-gold*).

La figure 6.2 présente la distribution du nombre de lexèmes en fonction du nombre de sens. Le niveau de polysémie, i.e. le nombre moyen de sens par lexème, est de 1,61 pour les noms et 2,12 pour les verbes.

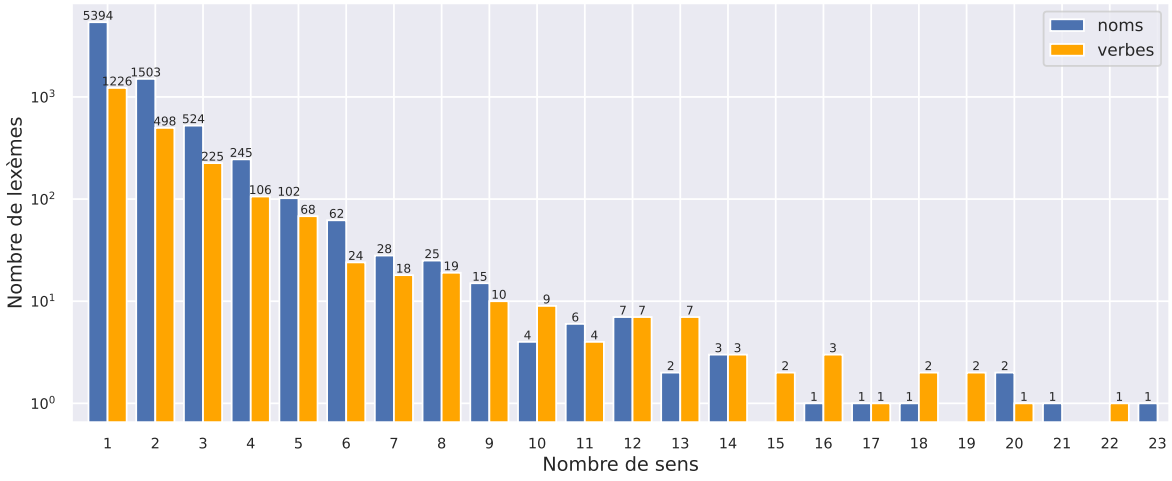


FIGURE 6.2 – Distribution du nombre de lexèmes en fonction du nombre de sens dans le BEL-RL-fr.

Les données sont ensuite divisées en échantillons d'entraînement, de validation et de test. Pour maintenir une cohérence dans les exemples durant la phase d'entraînement, plusieurs règles ont été appliquées pour la division des données :

1. Les exemples contenant des mots monosémiques sont placés dans l'échantillon d'entraînement, de manière à ce que les échantillons d'évaluation (validation ou test) ne contiennent pas ces cas triviaux où un seul sens est possible.
2. Tout lexème dans l'échantillon de test doit également apparaître dans l'échantillon d'entraînement, afin que le modèle ait vu chaque lexème au moins une fois pendant la phase d'entraînement.
3. Bien qu'un lexème soit connu, il est possible que le modèle n'ait pas vu tous ses sens.
4. La distribution des sens est gardée similaire dans chaque échantillon à celle du jeu de données entier, ce qui évite de biaiser le modèle vers des sens rares ou fréquents.
5. Si un exemple contient plus d'un sens à désambiguïser (par exemple à la fois un nom et un verbe), l'exemple n'est assigné qu'à un seul échantillon, pour éviter que le modèle ne

bénéficie d'une fuite de données.

Le tableau 6.3 présente la composition des échantillons d'entraînement, de validation et de test après application de la stratégie de division.

	Noms	**Verbes**
Nombre d'instances d'entraînement	13 645	5 484
Nombre d'instances de validation	2 660	874
Nombre d'instances de test	2 660	873

TABLE 6.3 – Nombre d'instances dans les échantillons d'entraînement, de validation et de test après division.

6.3 Systèmes de désambiguïsation lexicale

Pour évaluer la désambiguïsation lexicale des exemples lexicographiques, nous reprenons le modèle `EWISER` de Sinha et al. (2022) auquel nous ajoutons deux nouvelles configurations : l'intégration d'une matrice de poids aléatoires et d'une matrice de poids calculés à partir des plongements de graphe.

Nous couvrons différents types de systèmes de désambiguïsation lexicale :

— des *baselines* basées sur des heuristiques

— une *baseline* neuronale supervisée avec un perceptron multicouche

— le modèle `EWISER` et ses quatre configurations

— quatre grands modèles de langue génératifs

6.3.1 *Baselines*

Pour comparer les systèmes, nous appliquons trois *baselines* basées sur des heuristiques, une *baseline* aléatoire, la *baseline* classique du sens le plus fréquent et une *baseline* basée sur le voisinage lexical des exemples et des sens dans le `RL-fr`.

Nous entraînons également un perceptron multicouche qui constitue une *baseline* neuronale supervisée.

6.3.1.1 *Baselines* aléatoire et du sens le plus fréquent

Une première *baseline* simple consiste à choisir un sens aléatoirement parmi les sens possibles $\mathcal{S}(w)$ d'un lexème-cible w d'une instance (*Random Sense*). La performance de cette *baseline* doit être proche de l'exactitude théorique exprimée par :

$$\frac{1}{\#total_instances} \sum_{w} \frac{1}{\#\mathcal{S}(w)}$$

La *baseline* du sens le plus fréquent, *Most Frequent Sense* (`MFS`), est une *baseline* classique dans la tâche de désambiguïsation lexicale (cf. Section 5.6.3). Nous calculons les fréquences d'apparition des sens des noms et des verbes dans les échantillons d'entraînement (cf. Figure 6.3).

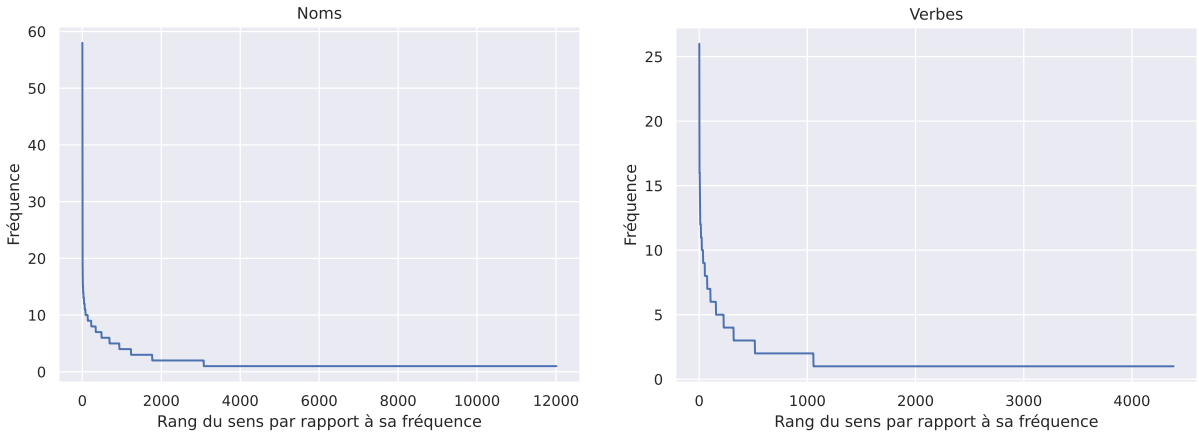


FIGURE 6.3 – Distribution des sens selon leur fréquence dans l'échantillon d'entraînement pour les noms (à gauche) et pour les verbes (à droite).

Nous vérifions si la distribution des sens suit une loi de Zipf selon la méthode décrite par Clauset et al. (2009) et implémentée dans une bibliothèque Python (Alstott et al., 2014). Le test statistique repose sur un estimateur de maximum de vraisemblance qui calcule le degré d'ajustement du modèle aux données observées, noté R, selon une valeur-p. Une valeur-p supérieure à 0,1 et un R positif signifie que la distribution suit une loi de Zipf (Clauset et al., 2009).

Pour les noms, la valeur R est de -39,6 avec une valeur-p de $1,17 \times 10^{-90}$. Pour les verbes, R est de -10,6 avec une valeur-p de $6,27 \times 10^{-8}$.

Nous pouvons conclure que les distributions des sens dans les échantillons d'entraînement pour les noms et pour les verbes ne suivent pas une loi de Zipf, autrement dit, la distribution des sens du `BEL-RL-fr` ne reflète pas celle d'un texte naturel.

Nous calculons toutefois les résultats du `MFS` à titre de comparaison. Dans le cas où plusieurs sens possibles présentent la fréquence maximale, nous choisissons un sens aléatoirement.

6.3.1.2 *Baseline* basée sur le voisinage lexical

Une autre *baseline* consiste à comparer le voisinage lexical d'un sens dans son contexte et dans le réseau lexico-sémantique. Le principe est similaire à l'algorithme Lesk qui compare les mots du contexte dans lequel se trouve le sens à désambiguïser et les mots présents dans les définitions de chaque sens possible (Lesk, 1986).

Le RL-fr ne présentant pas de définitions, nous utilisons le voisinage lexical direct des sens possibles dans le réseau, c'est-à-dire les nœuds reliés à chaque sens avec une distance de 1. On calcule ensuite un indice de Jaccard entre les lexèmes présents dans l'exemple et les voisins de chaque sens du lexème-cible dans le réseau.

Soit l'exemple suivant :

« Le long du fleuve, en été, toute sa famille se promène en `bicyclette`. »
où `bicyclette` est le lexème-cible à désambiguïser.

Il existe deux sens possibles pour `bicyclette` dans le `RL-fr` :
— node-26492 : `bicyclette I` → sens véhicule,
— node-41150 : `bicyclette II` → sens activité.

Le tableau 6.4 présente le voisinage de chaque sens dans le `RL-fr` et l'indice de Jaccard par rapport au voisinage lexical de l'exemple.

	Voisinage lexical	**Jaccard**
Exemple	["le","long","de", "fleuve", "en", "été, ",", "tout", "son", "famille", "se", "promener", "."]	-
node-26492 : `bicyclette I`	['cycliste', 'véhicule roulant', 'enfourcher', 'bécane', 'circuler', 'cyclotouriste', 'biclo', 'deux-roues', 'à', 'déplacer', 'guidon', 'en', 'cycle', 'descendre', 'monter', 'avoir l'air d'un crapaud sur une boîte d'allumettes', 'petite reine', 'rouler', 'bicyclette', 'escalader', 'escalade', 'grimper', 'pédaler', 'clou', 'vélocipède', 'biclou', 'vélo', 'faire']	0,025
node-41150 : `bicyclette II`	['bicyclette', 'faire', 'cyclisme', 'vélo', 'cycliste', 'pratiquer', 'petite reine', 'loisir', 'cyclable']	0

TABLE 6.4 – Exemple de la *baseline* basée sur le voisinage lexical.

Le sens avec l'indice de Jaccard le plus élevé constitue la prédiction, en l'occurrence `bicyclette I`, qui correspond au sens correct dans cet exemple. Dans le cas où plusieurs sens présentent le même indice de Jaccard maximal, nous sélectionnons un sens aléatoirement.

6.3.1.3 *Baseline* perceptron multicouche (MLP)

Outre des *baselines* heuristiques, nous présentons une *baseline* neuronale supervisée composée d'un perceptron multicouche.

Soit une phrase $x = (w_1, \ldots, w_n)$ contenant un lexème-cible w_t. La phrase est tokenisée et encodée par un modèle `CamemBERT` (Martin et al., 2020) pré-entraîné et gelé :

$$(\boldsymbol{h}_1, \ldots, \boldsymbol{h}_n) = \mathrm{BERT}(x), \quad \boldsymbol{h}_j \in \mathbb{R}^d$$

où d est la dimension de l'espace des plongements.

Le plongement $\boldsymbol{h}_t$ du lexème-cible est extrait, puis passé dans un perceptron à deux couches de paramètres W_1, b_1 et W_2, b_2 :

$$\begin{aligned}
\boldsymbol{u} &= W_1 \boldsymbol{h}_t + \boldsymbol{b}_1 \\
\boldsymbol{a} &= \mathrm{ReLU}(\boldsymbol{u}) \\
\tilde{\boldsymbol{a}} &= \mathrm{Dropout}(\boldsymbol{a}) \\
\boldsymbol{z} &= W_2 \tilde{\boldsymbol{a}} + \boldsymbol{b}_2
\end{aligned}$$

où $\boldsymbol{z} \in \mathbb{R}^{|\mathcal{V}|}$ est le vecteur de scores associé aux sens de $\mathcal{V}$. On utilise la fonction d'activation Unité Linéaire Rectifiée (ReLU) et un Dropout. Les probabilités sont ensuite obtenues avec une fonction softmax :

$$p(s \mid w_t, x) = \text{softmax}(\boldsymbol{z})_s$$

6.3.2 Modèle EWISER et intégration de plongements de graphe

Initialement, EWISER est un modèle neuronal supervisé pour la désambiguïsation lexicale enrichi par l'intégration d'informations sémantiques, d'une part de manière implicite *via* des plongements, et d'autre part de manière explicite *via* une matrice d'adjacence encodant des informations provenant du graphe (Bevilacqua and Navigli, 2020).

Sinha et al. (2022) proposent une adaptation du modèle pour la désambiguïsation du BEL-RL-fr, où seule la matrice d'adjacence est exploitée avec différentes configurations, leur objectif étant d'évaluer l'influence des informations structurelles du graphe (cf. Section 6.3.2.2). Notre contribution consiste à proposer une nouvelle configuration intégrant des plongements de graphe du RL-fr générés dans le chapitre 4.

6.3.2.1 Architecture centrale d'EWISER

Le cœur du modèle EWISER suit la même architecture proposée dans la *baseline* MLP mais fait intervenir d'autres composants.

Le plongement du lexème-cible $\boldsymbol{h}_t$ est obtenu avec CamemBERT de la même manière que dans la *baseline* MLP et est projeté dans l'espace des sens. On applique ensuite une BatchNorm, une couche de réseau de neurones de normalisation qui permet un entraînement plus stable et rapide (Luo et al., 2019).

Bevilacqua and Navigli (2020) choisissent également d'utiliser la fonction d'activation swish (Ramachandran et al., 2018), cette dernière démontrant de meilleures performances sur différentes tâches de TAL (Eger et al., 2018) :

$$\begin{aligned}\hat{\boldsymbol{h}} &= \text{BatchNorm}(\boldsymbol{h}_t)\\ \boldsymbol{u} &= W\hat{\boldsymbol{h}} + \boldsymbol{b}\\ \boldsymbol{h}' &= \text{swish}(\boldsymbol{u})\end{aligned}$$

où W et b sont des paramètres du modèle et $\text{swish}(x) = x \cdot \sigma(x)$, σ pour la fonction sigmoïde.

Les plongements obtenus sont ensuite calculés à partir d'une matrice de poids $O \in \mathbb{R}^{d \times |\mathcal{V}|}$ et les probabilités sont obtenues avec une fonction softmax :

$$\begin{aligned}\boldsymbol{z} &= \boldsymbol{h}'O\\ p(s \mid w_t, x) &= \text{softmax}(\boldsymbol{z})_s\end{aligned}$$

Dans Bevilacqua and Navigli (2020), la matrice O encode des plongements de sens disponibles tels que `LMMS 2048d` (Loureiro and Jorge, 2019) ou `SenEmBERT` (Scarlini et al., 2020). Toutefois, Sinha et al. (2022) n'exploitent pas cette composante du modèle et se concentrent uniquement sur l'intégration de connaissances provenant du graphe.

6.3.2.2 Intégration de plongements de graphe

Pour intégrer des informations provenant du graphe, `EWISER` présente une matrice d'adjacence A permettant d'encoder des poids sur les relations entre les sens.

Formellement, on considère un réseau lexico-sémantique : $G = \langle \mathcal{V}, \{R_{ij}, (i,j) \in \mathcal{V} \times \mathcal{V}\}, \omega \rangle$. $\mathcal{V}$ est l'ensemble des sens (nœuds). Pour deux sens i et j, R_{ij} est l'ensemble des relations entre i et j. Chacune de ces relations r possède un poids donné par une fonction $\omega(r) \in \mathbb{R}$, ainsi qu'un type encodant une fonction lexicale spécifique.

Nous pouvons construire une matrice d'adjacence A associée à G de la manière suivante :

$$A \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}, \forall (i,j) \in \mathcal{V} \times \mathcal{V}, A_{ij} = \sum_{r \in R_{ij}} \omega(r)$$

Nous reprenons les deux stratégies de calcul des poids `STRUCT` et `SEM` proposées par Sinha et al. (2022) et nous ajoutons deux nouvelles stratégies, `COS` et `RANDOM` :

— `STRUCT` : on considère le **nombre de relations** entre deux sens.

$$\omega(r) = 1, \text{ donc } \forall (i,j) \in \mathcal{V} \times \mathcal{V}, A_{ij} = |R_{ij}|$$

— `SEM` : Une fonction lexicale (relation) entre deux unités lexicales présente un **poids sémantique** dans le `RL-fr`. Dans la stratégie `SEM`, le poids $\omega(r)$ est défini par :

$$\omega(r) = 1 + s_r, \text{ où } s_r \in \{0, 1, 2\} \text{ est le poids sémantique de la relation } r.$$

— `COS` : on considère la **similarité cosinus entre les plongements de graphe $\boldsymbol{\theta_i}, \boldsymbol{\theta_j}$** de deux sens i, j.

$$A_{ij} = cos_{sim}(\boldsymbol{\theta_i}, \boldsymbol{\theta_j})$$

— `RANDOM` : on fixe **aléatoirement** le poids entre deux sens.

Dans la stratégie `COS`, nous utilisons les plongements de graphe générés pour chaque sens du `RL-fr` dans le chapitre 4. Nous testons l'hypothèse que le poids de la relation entre deux sens peut être représenté par la distance entre ses plongements de graphe respectifs.

La stratégie `RANDOM` permet de comparer l'influence des valeurs des poids des différentes stratégies.

On intègre ensuite la matrice d'adjacence A dans le modèle décrit précédemment en section 6.3.2.1 :

$$\hat{\boldsymbol{h}} = \text{BatchNorm}(\boldsymbol{h}_t)$$
$$\boldsymbol{u} = W\hat{\boldsymbol{h}} + \boldsymbol{b}$$
$$\boldsymbol{h}' = \text{swish}(\boldsymbol{u})$$
$$\boldsymbol{q} = \boldsymbol{h}'A + \boldsymbol{h}'$$

Enfin, les probabilités sur les sens sont obtenues par :

$$p(s \mid w_t, x) = \text{softmax}(\boldsymbol{q})_s$$

Une fois la matrice A intégrée, il est possible de garder les poids « gelés » ou de les entraîner. Les configurations avec la matrice entraînée sont notées avec un astérisque. Le tableau 6.5 fournit les notations des modèles que nous utiliserons dans nos expériences.

Modèle	**Spécificités**
`EWISER-A`	Architecture centrale d'`EWISER` sans matrice A
`EWISER+A RANDOM`	Poids aléatoires gelés dans la matrice A
`EWISER+A RANDOM*`	Poids aléatoires entraînés
`EWISER+A STRUCT`	Poids = nombre de relations entre les sens
`EWISER+A STRUCT*`	Stratégie `STRUCT` et poids entraînés
`EWISER+A SEM`	Poids = poids sémantique des relations dans le `RL-fr`
`EWISER+A SEM*`	Stratégie `SEM` et poids entraînés
`EWISER+A COS`	Poids = similarité cosinus entre les plongements de graphes des sens
`EWISER+A COS*`	Stratégie `COS` et poids entraînés

Table 6.5 – Configurations du modèle `EWISER` selon les informations relatives au graphe.

6.3.3 Grands modèles de langue

Comme nous l'avons décrit dans la section 5.4.3, il existe deux manières de traiter la tâche de désambiguïsation en *zero-shot* avec les grands modèles de langues :

1. sous la forme d'une tâche de classification : le modèle choisit le sens correct parmi une liste de sens possibles ;
2. sous la forme d'une tâche de génération de définition : le modèle génère une définition d'un lexème-cible à désambiguïser. La définition générée est comparée à une définition *gold*.

La première stratégie conserve le format de la tâche de classification et permet une évaluation contrôlée dans la mesure où il suffit de comparer la prédiction du modèle au *gold*. La seconde stratégie exploite davantage les capacités de génération des grands modèles de langue. Toutefois, l'évaluation est plus difficile car elle implique de trouver une manière de mesurer la similarité entre une définition générée et une définition de référence. Pour ce faire, les métriques utilisées sont généralement celles de la traduction automatique (Basile et al., 2025; Meconi et al., 2025).

Contrairement à WordNet, le `RL-fr` ne propose pas de définition pour toutes les unités lexicales, ce qui ne nous permet pas de comparer des définitions générées avec des *gold*. Nous considérons donc la tâche comme une tâche de classification.

Stratégie de *prompting*. Dans le *prompt*, une phrase avec un lexème-cible à désambiguïser est donnée en entrée à un grand modèle de langue. Les sens possibles du lexème-cible sont également fournis, décrits de deux manières différentes (cf. Tableau 6.6) :

1. par le voisinage lexical direct dans le réseau lexical `RL-fr`, c'est-à-dire l'ensemble des termes liés au sens d'un pas de 1 dans le réseau, toutes relations confondues ;
2. par un exemple d'utilisation du sens provenant du `BEL-RL-fr`.

Phrase d'entrée	Le long du fleuve, en été, toute sa famille se promène en `bicyclette`.
Description 1 : `RL-fr` voisinage	Les candidats sont : 1) ['cycliste', 'véhicule roulant', 'enfourcher', 'bécane', 'circuler', 'cyclotouriste', 'biclo', 'deux-roues', 'à', 'déplacer', 'guidon', 'en', 'cycle', 'descendre', 'monter', 'avoir l'air d'un crapaud sur une boîte d'allumettes', 'petite reine', 'rouler', 'bicyclette', 'escalader', 'escalade', 'grimper', 'pédaler', 'clou', 'vélocipède', 'biclou', 'vélo', 'faire', 'cyclotouriste'] 2) ['bicyclette', 'faire', 'cyclisme', 'vélo', 'cycliste', 'pratiquer', 'petite reine', 'loisir', 'cyclable']
Description 2 : `BEL-RL-fr` exemples	Les candidats sont : 1) La bicyclette était à ma taille, elle avait une lampe ronde et une dynamo, j'en caressais sans cesse le volume, la consistance, et je n'arrivais pas à croire qu'elle était à moi, que je pouvais avoir de tels objets « en vrai ». 2) L'agglomération de Nancy lance à son tour un système de vélos en libre service. Petite innovation par rapport aux initiatives précédentes en la matière : une Maison du vélo sera prochainement ouverte afin de promouvoir la pratique de la bicyclette

TABLE 6.6 – Deux types de description des sens possibles du lexème `bicyclette`.

Des travaux ont démontré la forte sensibilité des modèles à la manière de formuler les *prompts*, donnant lieu à des écarts de performances selon les stratégies adoptées (Mu et al., 2024; Errica et al., 2025). Shen et al. (2024) démontrent également que la langue du *prompt* peut influencer les sorties des modèles.

Pour la tâche de désambiguïsation lexicale, Meconi et al. (2025) appliquent 20 stratégies de *prompting* sur le jeu de données en anglais de SemEval-2007 : formulations explicites, par question-réponse ou dans un style conversationnel, génération de synonymes, par élimination de formulation. Les résultats comparent les performances de six grands modèles de langue [57] et

57. Les modèles utilisés sont : `Llama-3.2-1B.it`, `Phi-3.5-mini-4B-it`, `Phi-3-small-128k-7B-it`, `Gemma-2-9b-it`, `DeepSeek-V3` et `GPT-4o`.

montrent un écart de 5 points en moyenne entre le *prompt* le plus performant (P2), avec une micro-F1 de 63,9 % et les moins performants (P13, P20) avec 58,6 % (cf. Tableau 6.7).

P2 : instructions explicites	Read the sentence : <sentence> Choose the correct dictionary definition of the word "<word>" from the options : <definitions>. Provide as output only the correct dictionary definition
P13 : synonymes	In the sentence "<sentence>", the word "<word>" has a specific meaning. First, think about some synonyms that match its usage in this context. Then, choose the most appropriate definition from the following : <definitions>. Provide as output only the best definition
P20 : élimination	Consider the word "<word>" in the sentence : <sentence> Identify and exclude the incorrect definitions and then choose the best one : <definitions>. Provide as output only the best definition.

TABLE 6.7 – Stratégies de *prompting* adoptées dans Meconi et al. (2025).

Dans nos expériences, nous adoptons un *prompt* qui détaille la tâche que le modèle doit effectuer, ainsi que le format de la sortie attendue (json). Cela permet d'obtenir une sortie davantage contrôlée et simple à traiter. Nous utilisons également un *prompt* demandant au modèle de générer une définition du lexème à désambiguïser selon son contexte (**def**). La définition générée n'est pas évaluée en l'absence de définition *gold*, nous cherchons à observer si cette étape intermédiaire permet d'améliorer les performances du modèle.

Nous appliquons ensuite ces *prompts* avec les deux descriptions des sens possibles : le voisinage lexical dans le `RL-fr` (**lexical**) et un exemple d'utilisation du `BEL-RL-fr`, issu de l'échantillon d'entraînement (**usage**). Ces *prompts* sont formulés en français et en anglais, ce qui donne au total huit prompts différents. La figure 6.4 présente le *prompt* en français avec les sens décrits par leur voisinage lexical (**fr_lexical**).

```
Vous effectuez une tâche de désambiguïsation lexicale (WSD)
sur des exemples lexicographiques français.

Vous recevrez:
- Une phrase contenant le **lexème-cible** dans son contexte.
- Une liste de **sens** possibles pour ce lexème.
Chaque sens comprend :
- Une liste de voisins lexicaux (mots apparentés) provenant d'un graphe sémantique.
Votre tâche :
- Déterminer quel sens correspond le mieux à la signification du lexème-cible dans la phrase donnée.
- Ne produisez **que** le nombre du sens correct (pas d'explication, pas de texte) dans une
sortie json qui contient les clés suivantes :  sentence_id, target_word, choice.
Utilisez le format suivant :
"sentence_id":
"target_word":
"choice":
```

FIGURE 6.4 – *Prompt* en français pour la désambiguïsation lexicale. Les sens sont décrits avec le voisinage lexical.

Pour les verbes, nous adoptons uniquement la stratégie lexicale, le `BEL-RL-fr` ne fournissant pas d'exemples lexicographiques pour tous les sens candidats.

Modèles sélectionnés. Nous choisissons de tester quatre grands modèles de langue *open-weight*, accessibles *via* la plate-forme Ollama [58]. Nous fournissons, dans la mesure du possible, des détails sur les langues supportées et les types de données d'entraînement.

— `Gemma3-4b` (Team et al., 2025), modèle multimodal couvrant 140 langues ;

— `Qwen2.5-7b` (Qwen et al., 2025), entraîné sur du texte et couvrant 29 langues ;

— `Llama3.1-8b` (Grattafiori et al., 2024), modèle multilingue entraîné sur du texte (anglais, allemand, français, italien, portugais, hindi, espagnol et thaï) ;

— `Mistral-7b` (Jiang et al., 2023), entraîné sur du texte [59].

Pour des raisons de ressources matérielles disponibles, nous testons uniquement les versions avec le moins de paramètres.

6.4 Évaluer au-delà de la classification

Dans le chapitre 5, nous avons consacré une section sur l'évaluation classique pour la tâche de désambiguïsation et les biais liés aux données et aux métriques. Dans nos expériences, nous proposons d'évaluer les prédictions des modèles de manière plus détaillée, tout d'abord en utilisant plusieurs métriques, puis en exploitant les distributions de probabilités des modèles.

6.4.1 Métriques

La métrique traditionnelle pour évaluer la désambiguïsation lexicale sous forme de tâche de classification est la micro-F1, utilisée dans la plupart des travaux. Nous avons toutefois détaillé dans la section 5.6 du chapitre 5, qu'une micro-F1 peut varier selon la fréquence d'apparition de certains sens. Un sens très fréquent et facilement prédit va tirer la micro-F1 vers le haut, masquant ainsi la performance du système sur des sens peu fréquents.

Le `BEL-RL-fr` n'étant pas représentatif de la distribution réelle des sens dans un texte naturel, le déséquilibre dans la fréquence des sens est moins marqué dans nos expériences. Nous évaluerons toutefois les performances des modèles avec une **micro-F1** et une **macro-F1**.

6.4.2 Évaluation restreinte *vs* Évaluation globale

Afin d'analyser les prédictions d'un système supervisé de désambiguïsation lexicale, nous distinguons deux configurations d'évaluation :

— une évaluation **restreinte** : le sens prédit est le sens dont la probabilité est la plus haute **parmi les sens possibles du lexème-cible**. Cette configuration est celle traditionnellement appliquée dans les travaux sur la désambiguïsation lexicale.

— une évaluation **globale** : le sens prédit est le sens dont la probabilité est la plus haute **parmi tout le vocabulaire**.

58. `https://ollama.com/`.

59. L'information sur les langues supportées n'est indiquée ni le site officiel, ni le rapport technique..

On rappelle que la composante centrale d'un système supervisé de désambiguïsation lexicale est l'estimation de la probabilité conditionnelle :

$$P(s_i \mid c)$$

où s_i est un sens du lexème-cible et c, le contexte d'apparition.

Les systèmes neuronaux supervisés calculent un score pour chaque sens possible :

$$\text{score}(s_i, c) = f_\theta(s_i, c)$$

où f_θ est une fonction neuronale paramétrée, généralement basée sur des modèles de langue contextuels. Ces scores sont ensuite convertis en une distribution de probabilité avec une fonction softmax :

$$P(s_i \mid c) = \frac{e^{\text{score}(s_i,c)}}{\sum_{j=1}^{n} e^{\text{score}(s_j,c)}}$$

Une fois la distribution de probabilités obtenue, le système choisit le sens avec la probabilité la plus haute.

En évaluation *restreinte*, la prédiction finale correspond au sens avec la probabilité la plus haute **parmi les sens possibles du lexème-cible**, $\mathcal{S}(w)$:

$$\hat{s}_{\mathcal{S}(w)} = \arg\max_{s_i \in \mathcal{S}(w)} P(s_i \mid c)$$

Ce scénario est la configuration classique en désambiguïsation lexicale, l'objectif étant d'évaluer si le système attribue le sens correct d'un lexème-cible parmi ses sens possibles. On se concentre uniquement sur les probabilités du sous-ensemble des sens possibles.

En pratique, le système calcule la distribution de probabilités sur tout le vocabulaire, qui correspond à l'ensemble des sens possibles de tous les mots. Il est alors possible de récupérer le sens avec la probabilité la plus haute **parmi tout le vocabulaire**, $\mathcal{V}$:

$$\hat{s}_{\mathcal{V}} = \arg\max_{s_i \in \mathcal{V}} P(s_i \mid c)$$

Cette configuration, que nous appellerons l'**évaluation globale**, permet d'observer si le système attribue une probabilité plus élevée à un sens qui n'est pas parmi les sens possibles. Si cette manière d'évaluer s'éloigne de l'objectif principal de la tâche de désambiguïsation lexicale, elle met davantage en valeur le comportement réel du modèle vis-à-vis des prédictions qu'il produit.

L'évaluation globale implique que les modèles produisent des scores sur tous les sens du vocabulaire. Cette configuration n'est pas pertinente sur les *baselines* aléatoire et `MFS` et les grands modèles de langue qui n'ont accès qu'aux sens possibles du lexème-cible. Bien qu'elle soit

applicable sur la *baseline* basée sur le voisinage lexical, nous ne testons ici que la configuration classique en évaluation restreinte.

Par conséquent, l'évaluation restreinte et l'évaluation globale sont appliquées uniquement sur les modèles supervisés, dont les résultats sont décrits en section 6.6.

6.5 QR 1 : performances des systèmes en évaluation restreinte

Dans cette section, nous répondrons à la première question de recherche qui s'intéresse aux performances de différents systèmes pour la désambiguïsation lexicale des exemples du `BEL-RL-fr`.

Les modèles sont évalués sur le même jeu de données de test avec des performances données en macro-F1 (M-F1) et micro-F1 (m-F1). On applique une configuration d'évaluation classique où le sens prédit correspond à celui dont la probabilité est la plus élevée **parmi les sens possibles du lexème-cible** (évaluation restreinte). Cela permet ainsi une comparaison directe entre tous les modèles, les *baselines* heuristiques et les grands modèles de langue ne pouvant pas être testés avec une évaluation globale.

Le tableau 6.8 présente les résultats sur les systèmes présentés en section 6.3.

		Noms		**Verbes**	
Type de modèle	**Modèle**	**M-F1**	**m-F1**	**M-F1**	**m-F1**
Baselines heuristiques	`Random Sense`	13,23	23,95	6,59	14,77
	`MFS`	10,99	24,09	6,34	15,92
	`Voisinage lexical`	20,04	25,51	13,97	14,89
Baseline ML	`MLP`	47,35	56,09	23,03	31,96
`EWISER`	`EWISER-A`	**52,38**	**61,17**	30,28	40,78
	`EWISER+A RANDOM`	48,50	57,29	29,94	40,55
	`EWISER+A RANDOM*`	48,03	57,52	29,70	39,97
	`EWISER+A STRUCT`	49,57	57,29	33,16	41,47
	`EWISER+A STRUCT*`	51,07	58,46	31,38	40,32
	`EWISER+A SEM`	46,14	53,68	31,74	40,55
	`EWISER+A SEM*`	47,14	54,66	31,86	40,32
	`EWISER+A COS`	50,21	58,23	**33,27**	**43,76**
	`EWISER+A COS*`	50,86	58,76	32,74	43,30
`LLM`	`Qwen2.5-7b`	30,82	29,70	26,87	23,25
	`Gemma3-4b`	28,11	32,97	20,62	23,37
	`Llama3.1-8b`	28,00	33,23	23,42	19,82
	`Mistral-7b`	32,58	33,92	17,74	19,36

TABLE 6.8 – Performance (%) en macro-F1 (M-F1) et micro-F1 (m-F1) des modèles de désambiguïsation lexicale sur `BEL-RL-fr` sur les noms et les verbes en évaluation restreinte. Les meilleurs résultats pour chaque configuration sont présentés en gras.

6.5.1 Meilleures performances avec un modèle neuronal supervisé

Dans un premier temps, nous observons que les performances sont meilleures sur les noms communs que sur les verbes, ce qui est en accord avec les résultats de Sinha et al. (2022).

Par ailleurs, les modèles les plus performants sont les mêmes quelque soit la métrique considérée. Toutefois, nous notons une différence notable d'environ 10 points entre les valeurs de micro-F1 et de macro-F1 sur les modèles les plus performants. Nous détaillons cette observation dans la section 6.5.2.

Modèles `EWISER`. Tant pour les noms que pour les verbes, le système présentant les meilleures résultats est un modèle neuronal supervisé, `EWISER-A` pour les noms avec une micro-F1 de 61,17 % et `EWISER+A COS` pour les verbes avec une micro-F1 de 43,76 %. Nous notons que l'ajout de la matrice A ne permet pas d'améliorer les performances sur les noms, quelle que soit la stratégie adoptée. Cependant, les performances sont meilleures avec `EWISER+A COS` pour les verbes, avec une micro-F1 passant de 40,78 % à 43,30 %. L'évaluation détaillée de l'influence de l'intégration de la matrice est présentée dans la section 6.6.3.

Baselines. Concernant les *baselines*, la valeur de la *baseline* aléatoire est une moyenne de cinq tirages. Les résultats obtenus en micro-F1 sont logiquement proches des exactitudes théoriques qui sont de 23,90 % pour les noms et 14,58 % pour les verbes.

Nous remarquons ensuite que la *baseline* du sens le plus fréquent (`MFS`) est proche de la *baseline* aléatoire, ce qui confirme qu'elle n'est pas adaptée lorsque la distribution des sens ne suit pas une loi de Zipf.

La *baseline* basée sur le voisinage lexical est également peu performante, ses scores étant proches de la *baseline* aléatoire. Les valeurs des indices de Jaccard sont relativement faibles, l'indice maximal est de 0,25 pour les noms et 0,09 pour les verbes, ce qui montre une intersection limitée entre le voisinage lexical des sens et les lexèmes du contexte.

Grands modèles de langue. Dans le tableau 6.8, nous présentons les meilleurs résultats en micro-F1 pour chaque grand modèle de langue. Les stratégies de *prompting* correspondantes sont données dans le tableau 6.9.

Modèle	Noms	Verbes
`Qwen2.5-7b`	`en_lexical`	`en_lexical_def`
`Gemma3-4b`	`fr_lexical_def`	`fr_lexical_def`
`Llama3.1-8b`	`fr_lexical_def`	`en_lexical_def`
`Mistral-7b`	`fr_usage`	`en_lexical`

TABLE 6.9 – Stratégie de *prompting* donnant les meilleurs résultats en micro-F1 selon les modèles.

Pour les noms, les résultats des grands modèles de langue sont plus ou moins constants à travers les modèles, avec une micro-F1 moyenne de 32 %. Pour les verbes, on observe un écart un peu plus marqué entre `Qwen2.5-7b` et `Gemma3-4b` autour de 23 % de micro-F1 et `Llama3.1-8b`

et `Mistral-7b` autour de 19 %. La description des sens par le voisinage lexical dans le `RL-fr` semble donner les meilleurs résultats sauf pour `Mistral-7b` sur les noms.

Les performances sont dans l'ensemble plus élevées par rapport aux *baselines* heuristiques mais bien moindres qu'avec les modèles `EWISER`. Le meilleur modèle sur les noms est `Mistral-7b` avec un *prompt* en français et qui décrit les sens possibles avec des exemples lexicographiques du `BEL-RL-fr` (`fr_usage`). Sur les verbes, `Qwen2.5-7b` présente la macro-F1 la plus élevée avec 26,87 % mais a une micro-F1 de 23,25 % qui est légèrement plus basse que celle de `Gemma3-4b`, à 23,37 %.

Nous proposons une analyse approfondie des résultats des grands modèles de langue en section 6.8.

6.5.2 Des données d'évaluation qui influencent les métriques

Dans cette section, nous détaillons les raisons de l'écart de valeurs entre la macro-F1 et la micro-F1. L'analyse se base sur les résultats des deux modèles les plus performants sur les noms et les verbes, respectivement `EWISER-A` et `EWISER+A COS`.

Pour les noms, `EWISER-A` présente une macro-F1 de 52,38 % et une micro-F1 de 61,17 %, ce qui représente un écart de 8,79 points. Comme nous l'expliquons dans la section 5.5 du chapitre 5, la micro-F1 donne le même poids à chaque instance, contrairement à la macro-F1 qui donne le même poids à chaque classe.

Ainsi nous calculons la fréquence d'apparition des sens, c'est-à-dire le nombre d'instances par sens dans le jeu de données de test (cf. Figure 6.5). Pour les noms, il existe 1 658 sens dans le jeu de données de test, composé de 2 660 instances. On compte 1 044 sens n'apparaissant que dans une seule instance, soit près de 63 %. Par ailleurs, sur ces 1 044 instances, 54 % sont incorrectement prédites par le modèle, ce qui tire considérablement la macro-F1 vers le bas. En revanche, ces sens mono-instance incorrectement prédits auront moins de poids dans un calcul de micro-F1.

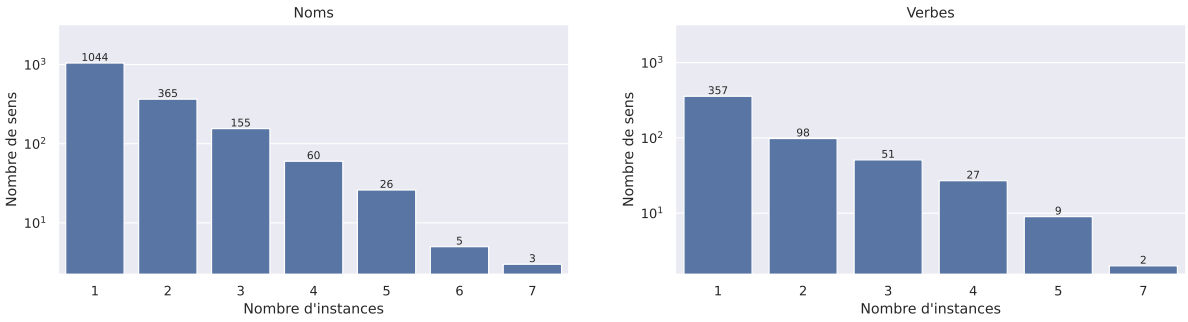


FIGURE 6.5 – Distribution du nombre d'instances par sens dans l'échantillon de test pour les noms et les verbes.

Nous observons une tendance similaire sur les verbes avec le modèle le plus performant, `EWISER+A COS`, présentant une macro-F1 de 33,27 % et une micro-F1 de 43,76 %, ce qui représente un écart de 10,49 points.

Dans le jeu de données de test composé de 873 instances, 544 sens sont présents et 357 (66 %) d'entre eux n'apparaissent que dans une seule instance (cf. Figure 6.5). Sur ces 357 instances,

71 % sont incorrectement prédites par le modèle.

6.5.3 Conclusion de la QR 1

La première question de recherche s'intéressait à déterminer le système le plus performant sur la tâche de désambiguïsation lexicale des exemples lexicographiques du `BEL-RL-fr` avec une évaluation classique.

Les résultats montrent que le modèle le plus performant sur les noms est `EWISER-A`, qui n'intègre pas de connaissances relatives au `RL-fr` mais qui dépasse les *baselines* heuristiques et un simple perceptron multicouche. Les grands modèles de langues présentent des résultats davantage proches de *baselines* que des modèles `EWISER` (cf. Section 6.8).

En revanche, l'ajout des plongements de graphe augmente les performances d'`EWISER` sur les verbes avec une micro-F1 passant de 40,78 % à 43,76 %. Nous détaillons ces résultats dans la deuxième question de recherche en section 6.6.

En accord avec les observations sur les biais du chapitre 5, nous remarquons un écart dans les valeurs des performances selon la métrique choisie. Avec une macro-F1, les performances perdent environ 9 points de pourcentage sur les meilleurs modèles par rapport à une micro-F1. Cette différence s'explique par le biais de la composition des données d'évaluation qui présentent une grande proportion de sens qui n'apparaissent que dans une seule instance et qui sont incorrectement prédits. La performance sur ces sens ne dépend ainsi que d'une seule prédiction du modèle, donnant une valeur extrême de F1 sur cette classe (égale à 1 ou 0). Ces cas de figure ont un fort poids lorsque les performances sont calculées avec une macro-F1.

6.6 QR 2 : influence de l'intégration de connaissances

On se propose dans cette section de répondre à la deuxième question de recherche qui se focalise sur l'influence de l'intégration de connaissances relatives au graphe, dont les plongements de graphe du `RL-fr` avec la stratégie `COS`. La matrice A encode des poids entre des sens liés dans le graphe, les différentes stratégies variant la valeur de ce poids.

Pour comprendre les résultats, nous présentons une analyse des prédictions des modèles pour déterminer l'influence de l'intégration de la matrice, notamment sur les probabilités des prédictions.

6.6.1 Comparaison des prédictions des modèles

Dans la première question de recherche (cf. Section 6.5), les résultats montrent que l'ajout de connaissances *via* la matrice d'adjacence A fait baisser les performances d'`EWISER` pour les noms. Nous analysons une instance correctement désambiguïsée par `EWISER-A` mais incorrectement désambiguïsée par la majorité des modèles `EWISER+A` :

« D'ailleurs les mamans non averties, qui cherchent en toute confiance des **chaussons** et un tablier de peinture pour une première rentrée en maternelle, après le 15 août, ne recommencent

jamais cette erreur de parent d'élève débutant ! » (cit6158)

Le lexème **chausson** présente trois acceptions possibles dans le `RL-fr` [60] :

— `chausson I.1` : chaussures d'intérieur,

— `chausson I.2` : chaussures pour une activité donnée,

— `chausson II` : aliment.

Le sens correct dans l'instance en exemple est `chausson I.1`. Le tableau 6.10 présente le top 3 des sens s_i classés par leurs probabilités $P(s_i)$ des modèles `EWISER`.

Modèle	s_1	$P(s_1)$	s_2	$P(s_2)$	s_3	$P(s_3)$
EWISER-A	chausson I.1	0,5183	chausson I.2	0,4397	chausson II	0,0381
EWISER+A RANDOM	serviette III	0,5058	chausson I.2	0,3559	chausson I.1	0,1285
EWISER+A RANDOM*	chausson I.2	0,6846	chausson I.1	0,2953	chaussure I	0,0068
EWISER+A STRUCT	chausson I.2	0,8706	chausson I.1	0,1294	chausson II	0,0000
EWISER+A STRUCT*	chausson I.1	0,5571	chausson I.2	0,4429	chaussure I	0,0000
EWISER+A SEM	chausson I.2	0,9634	chausson I.1	0,0366	chausson II	0,0000
EWISER+A SEM*	chausson I.2	0,7941	chausson I.1	0,2059	chausson II	0,0000
EWISER+A COS	chausson I.2	0,9999	chausson I.1	0,0000	chaussonnier.N	0,0000
EWISER+A COS*	chausson I.2	0,9885	chausson I.1	0,0108	serviette III	0,0003

TABLE 6.10 – Top 3 des sens s_i et des probabilités $P(s_i)$ des modèles `EWISER` sur l'instance cit6158 ayant pour *gold* = `chausson I.1`.

Dans un premier temps, nous remarquons qu'`EWISER-A` prédit le sens correct avec une probabilité maximale de 0,5138. Le sens `chausson I.2` est le deuxième sens au classement avec une probabilité de 0,4397 et enfin `chausson II` présente une probabilité plus faible à 0,0381. Nous notons que les probabilités sont distribuées majoritairement sur les sens possibles du lexème-cible.

Dans un second temps, sur les modèles intégrant la matrice d'adjacence A, six modèles sur huit prédisent incorrectement `chausson I.2` avec une probabilité relativement élevée, allant de 0,6846 à 0,9999. Seul `EWISER+A STRUCT*` prédit correctement le sens correct avec une probabilité de 0,5571, qui est légèrement plus élevée qu'`EWISER+A`. La distinction entre `chausson I.1` et `chausson I.2` est sémantiquement fine, ce qui explique la confusion des modèles à prédire le sens correct.

Certains modèles `EWISER+A` prédisent des sens qui ne font pas partie des sens possibles du lexème-cible. Par exemple, `EWISER+A RANDOM` prédit `serviette III` en première position avec une probabilité de 0,5058 et le sens correct apparaît en troisième position avec une probabilité de 0,1285. Nous remarquons également la présence de sens tels que `chaussure I` ou `chaussonnier.N`.

Nous pouvons supposer que les modèles intégrant la matrice sont influencés vers des sens reliés aux sens possibles du lexème-cible, se traduisant en une distribution des probabilités sur un ensemble de sens plus large que le sous-ensemble des sens possibles.

60. `https://spiderlex.atilf.fr/fr/q/*chausson***`

Pour les verbes, nous avons observé que le modèle `EWISER+A COS` présente les meilleures performances avec une micro-F1 de 43,76 % contre 40,78 % pour `EWISER-A`. L'ajout des plongements de graphe permet vraisemblablement d'augmenter les performances. Comme pour les noms, nous prenons l'exemple d'une instance correctement prédite par `EWISER+A COS` et nous comparons les prédictions des autres modèles :

« Le savon **lave**-t-il mieux que le gel douche ? » (cit19262)

Le verbe **laver** présente dix acceptions différentes dans le `RL-fr`. Dans cet exemple, le sens correct est `laver II`[61]. Les verbes présentent généralement un niveau de polysémie plus élevée que les noms communs, rendant leur désambiguïsation difficile.

Le top 3 des sens prédits par chaque modèle `EWISER` et les probabilités correspondantes sont détaillés dans le tableau 6.11.

`EWISER+A COS` prédit le sens correct avec la probabilité maximale de 0,4739, suivi du sens `nettoyer I.1b` avec une probabilité légèrement plus faible à 0,4282. Le sens correct n'est pas prédit avec un net écart de probabilités par rapport au second sens prédit, indiquant une proximité sémantique entre les deux sens, qui sont effectivement reliés par une relation de synonymie dans le `RL-fr`. Cela est confirmé par les résultats des autres modèles avec six sur huit modèles prédisant `nettoyer I.1b` en première position avec des probabilités allant de 0,1904 à 0,9982.

Outre `EWISER+A STRUCT` et `EWISER+A COS*` qui prédisent le sens correct avec la probabilité maximale, nous remarquons que trois modèles prédisent `laver II` en deuxième position et deux autres modèles, en troisième position. Sur ces cinq derniers modèles, avec une évaluation classique, on considère que le sens correct est bien prédit car celui-ci présente la probabilité maximale **parmi les sens possibles du lexème-cible**. Dans ce cadre d'évaluation, aucune différence n'est visible entre un sens prédit correctement avec une probabilité de 0,4739 (`EWISER+A COS`) et un sens prédit correctement avec une probabilité de 0,0005 (`EWISER+A SEM*`). Par conséquent, une évaluation basée simplement sur une métrique telle qu'une micro-F1 montre que sur cette instance, huit modèles sur neuf prédisent correctement le *gold* et un modèle se trompe.

Modèle	s_1	$P(s_1)$	s_2	$P(s_2)$	s_3	$P(s_3)$
`EWISER+A COS`	`laver II`	0,4739	`nettoyer I.1b`	0,4282	`laver V.1`	0,0236
`EWISER-A`	`nettoyer I.1b`	0,2947	`laver II`	0,1933	`laver I.1`	0,1137
`EWISER+A RANDOM`	`nettoyer I.1b`	0,3238	`laver II`	0,1455	`laver VI`	0,1042
`EWISER+A RANDOM*`	`nettoyer I.1b`	0,1904	`laver VI`	0,1598	`laver I.1`	0,1009
`EWISER+A STRUCT`	`laver II`	0,2913	`nettoyer I.1b`	0,2912	`laver IV`	0,0959
`EWISER+A STRUCT*`	`nettoyer I.1b`	0,5556	`laver II`	0,1936	`laver VI`	0,0820
`EWISER+A SEM`	`nettoyer I.1b`	0,8777	`doubler V`	0,1032	`laver II`	0,0144
`EWISER+A SEM*`	`nettoyer I.1b`	0,9982	`bouffer I.1a`	0,0006	`laver II`	0,0005
`EWISER+A COS*`	`laver II`	0,4474	`nettoyer I.1b`	0,4300	`laver I.1`	0,0333

TABLE 6.11 – Top 3 des sens s et des probabilités des modèles `EWISER` sur l'instance cit19262 ayant pour *gold* = `laver II`.

61. `https://spiderlex.atilf.fr/fr/q/*laver***`

Afin de mettre en évidence ce phénomène, nous introduisons une nouvelle configuration d'évaluation, que nous nommons **l'évaluation globale**, qui consiste à vérifier si le *gold* est prédit avec la probabilité maximale sur tout le vocabulaire (cf. Section 6.4.2).

6.6.2 Une baisse des scores en évaluation globale

Nous comparons les performances avec une évaluation globale, c'est-à-dire le sens prédit à la probabilité maximale sur tout le vocabulaire, notée $\hat{s}_{\mathcal{V}}$. En évaluation restreinte, le sens prédit présente la probabilité maximale sur les sens possibles du lexème-cible, notée $\hat{s}_{\mathcal{S}(w)}$.

On distingue ainsi quatre cas de figure, décrits par des exemples dans le tableau 6.12 :

1. gold $= \hat{s}_{\mathcal{V}} \wedge$ gold $= \hat{s}_{\mathcal{S}(w)}$: le sens *gold* est prédit correctement avec la probabilité la plus élevée sur tout le vocabulaire, donc sur les sens possibles du lexème-cible.
2. gold $\neq \hat{s}_{\mathcal{V}} \wedge$ gold $= \hat{s}_{\mathcal{S}(w)}$: le sens *gold* est prédit correctement mais ne présente pas la probabilité maximale sur tout le vocabulaire.
3. gold $\neq \hat{s}_{\mathcal{V}} \wedge \hat{s}_{\mathcal{V}} = \hat{s}_{\mathcal{S}(w)}$: le sens *gold* n'est pas prédit car le modèle a attribué la probabilité maximale à un autre sens du lexème-cible.
4. gold $\neq \hat{s}_{\mathcal{V}} \wedge \hat{s}_{\mathcal{V}} \neq \hat{s}_{\mathcal{S}(w)}$: le sens *gold* n'est pas prédit et le modèle a attribué la probabilité maximale à un sens qui n'est pas un sens possible du lexème-cible.

Cas	Inst	gold	$\hat{s}_{\mathcal{V}}$	P($\hat{s}_{\mathcal{V}}$)	$\hat{s}_{\mathcal{S}(w)}$	P($\hat{s}_{\mathcal{S}(w)}$)
1. gold $= \hat{s}_{\mathcal{V}} \wedge$ gold $= \hat{s}_{\mathcal{S}(w)}$	cit16234	`témoin.N.masc II.2`	`témoin.N.masc II.2`	0,99	`témoin.N.masc II.2`	0,99
2. gold $\neq \hat{s}_{\mathcal{V}} \wedge$ gold $= \hat{s}_{\mathcal{S}(w)}$	cit29886	`terrain II`	`champ 1 II.1`	0,4807	`terrain II`	0,0113
3. gold $\neq \hat{s}_{\mathcal{V}} \wedge \hat{s}_{\mathcal{V}} = \hat{s}_{\mathcal{S}(w)}$	cit11264	`mêmère II`	`mêmère I`	0,4918	`mêmère I`	0,4918
4. gold $\neq \hat{s}_{\mathcal{V}} \wedge \hat{s}_{\mathcal{V}} \neq \hat{s}_{\mathcal{S}(w)}$	cit26060	`navet II`	`documentaire.N`	0,3207	`navet I`	0,0006

TABLE 6.12 – Cas de figure des prédictions selon l'évaluation globale ou l'évaluation restreinte.

Les tableaux 6.13 et 6.14 présentent les performances du modèle `EWISER` avec et sans l'intégration de connaissances relatives au graphe, respectivement sur les noms et sur les verbes. Les performances sont données en macro-F1 et en micro-F1 sur les deux types d'évaluation, globale et restreinte.

Le scénario optimal serait d'obtenir une micro-F1 identique dans les deux évaluations, ce qui signifierait que tous les sens correctement prédits présentent la probabilité maximale sur tout le vocabulaire. Cependant, pour les noms comme pour les verbes, on remarque que la micro-F1 en évaluation globale est plus faible qu'en évaluation restreinte. Il existe donc des instances où les modèles ont prédit le sens correct mais ont attribué la probabilité maximale à des sens qui ne font pas partie des sens possibles du lexème (cas de figure n°2).

Pour les noms, l'écart entre la micro-F1 globale et la micro-F1 restreinte est plus marqué dans les modèles intégrant la matrice d'adjacence : en moyenne 8,22 points de différence (allant de 3 à 13,45 points) contre 1,96 points pour `EWISER-A`.

Pour les verbes, les écarts les plus faibles sont de 2,64 pour `EWISER+A RANDOM`, 2,75 pour `EWISER-A` et 2,98 pour `EWISER+A RANDOM*`. Les autres modèles présentent des écarts compris entre 3,67 et 10,54.

Les figures 6.6 et 6.7 présentent le nombre d'instances correctement prédites en évaluation restreinte sur les noms et les verbes. Nous faisons la distinction entre les cas où le *gold* présente

	Évaluation globale		Évaluation restreinte	
Modèle	**M-F1**	**m-F1**	**M-F1**	**m-F1**
EWISER-A	**49,91**	**59,21**	**52,38**	**61,17**
EWISER+A RANDOM	44,83	54,29	48,50	57,29
EWISER+A RANDOM*	41,62	51,92	48,03	57,52
EWISER+A STRUCT	35,24	44,02	49,57	57,29
EWISER+A STRUCT*	42,57	51,32	51,07	58,46
EWISER+A SEM	31,44	40,23	46,14	53,68
EWISER+A SEM*	33,42	41,92	47,14	54,66
EWISER+A COS	43,91	52,86	50,21	58,23
EWISER+A COS*	44,49	53,53	50,86	58,76

TABLE 6.13 – Performances (%) en macro-F1 (M-F1) et micro-F1 (m-F1) du modèle EWISER avec et sans intégration de connaissances relatives au graphe sur les noms. Les meilleurs résultats sont donnés en gras.

	Évaluation globale		Évaluation restreinte	
Modèle	**M-F1**	**m-F1**	**M-F1**	**m-F1**
EWISER-A	27,69	38,03	30,28	40,78
EWISER+A RANDOM	27,76	37,91	29,94	40,55
EWISER+A RANDOM*	26,45	36,99	29,70	39,97
EWISER+A STRUCT	22,46	30,93	33,16	41,47
EWISER+A STRUCT*	23,59	33,22	31,38	40,32
EWISER+A SEM	22,57	30,81	31,74	40,55
EWISER+A SEM*	22,24	31,04	31,86	40,32
EWISER+A COS	28,35	39,29	**33,27**	**43,76**
EWISER+A COS*	**28,93**	**39,63**	32,74	43,30

TABLE 6.14 – Performances (%) en macro-F1 (M-F1) et micro-F1 (m-F1) du modèle EWISER avec et sans intégration de connaissances relatives au graphe sur les verbes. Les meilleurs résultats sont donnés en gras.

la probabilité maximale sur tout le vocabulaire (cas de figure 1 en bleu) et seulement sur les sens possibles du lexème-cible (cas de figure 2 en orange).

Noms. Avec le modèle le plus performant, EWISER-A, il existe 52 instances où le sens correct est prédit mais ne présente pas la probabilité maximale. Ce nombre augmente sur les modèles qui intègrent la matrice A, notamment sur EWISER+A SEM, EWISER+A SEM*, et EWISER+A STRUCT avec plus de 300 instances. Ce phénomène explique ainsi la forte différence dans leur micro-F1 restreinte et globale (-13 points en moyenne).

Nous précisons que les 52 instances $gold \neq \hat{s}_{\mathcal{V}}$ de EWISER-A ne sont pas toutes incluses dans les instances $gold \neq \hat{s}_{\mathcal{V}}$ des autres modèles. Les modèles EWISER+A présentent en moyenne 30 instances communes avec EWISER-A. Seules deux instances sont communes à tous les modèles.

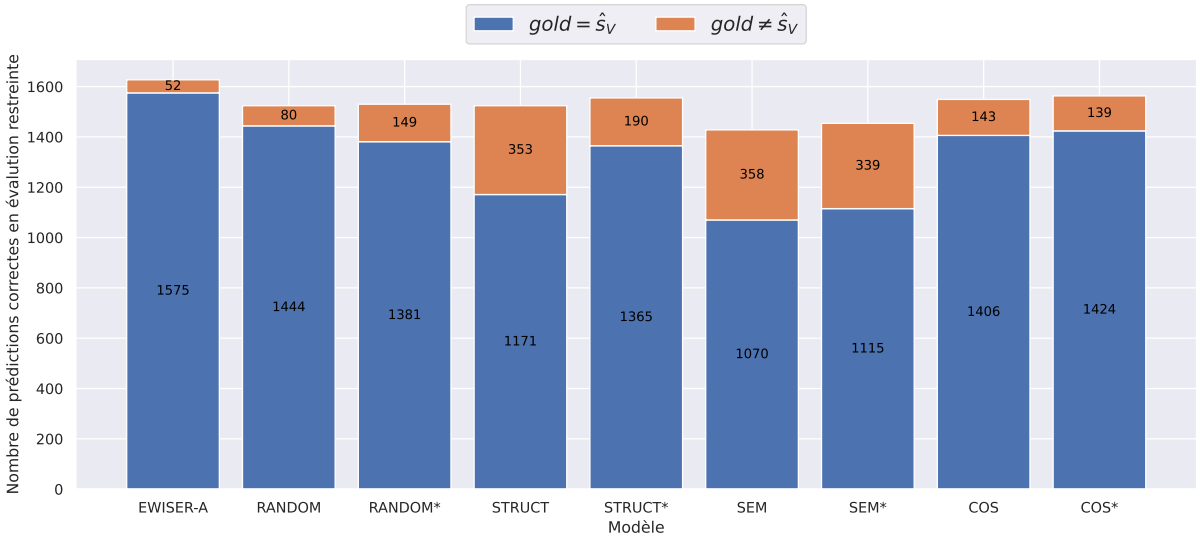


FIGURE 6.6 – Nombres d'instances correctement prédites en évaluation restreinte ($gold = \hat{s}_{\mathcal{S}(w)}$) selon les modèles `EWISER` sur les noms.

Verbes. Sur les verbes, les résultats sont différents dans la mesure où les deux modèles avec la stratégie `COS` présentent les meilleures performances : 43,76 % pour `EWISER+A COS` et 43,30 % pour `EWISER+A COS*` contre 40,78 % pour `EWISER-A` en évaluation restreinte.

Les modèles `COS` présentent davantage d'instances correctement prédites avec la probabilité maximale ($gold = \hat{s}_{\mathcal{V}}$) mais également d'instances correctement prédites sans la probabilité maximale ($gold \neq \hat{s}_{\mathcal{V}}$) par rapport à `EWISER-A`.

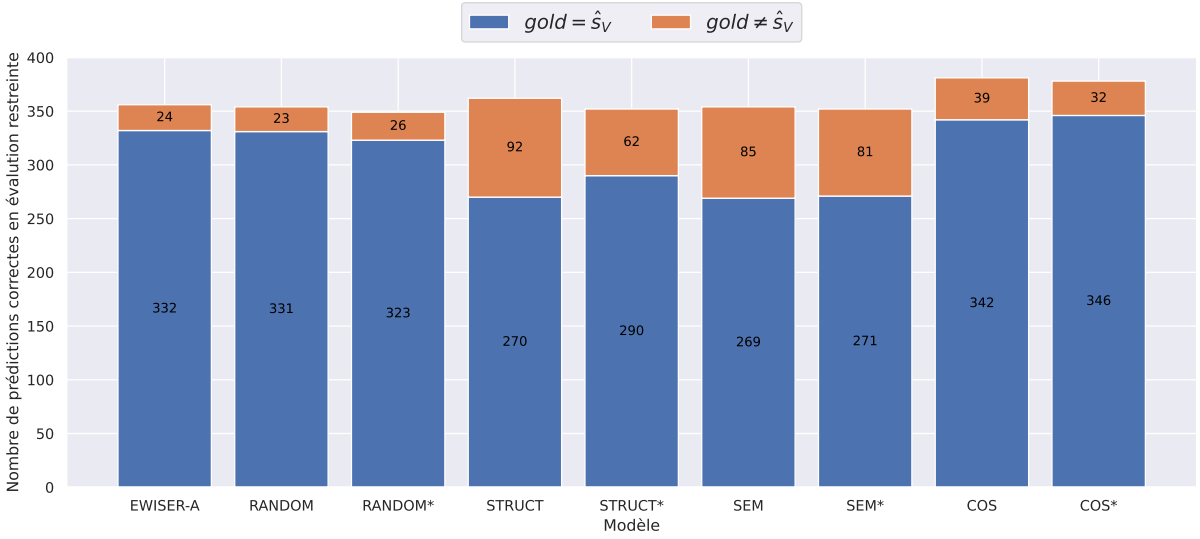


FIGURE 6.7 – Nombres d'instances correctement prédites en évaluation restreinte (gold = $\hat{s}_{\mathcal{S}(w)}$) selon les modèles `EWISER` sur les verbes.

Comme observé dans la section 6.5, nous supposons que l'ajout de la matrice influence la distribution de probabilités du système vers des sens liés aux sens possibles du lexème-cible, ce qui expliquerait l'augmentation des instances $gold \neq \hat{s}_{\mathcal{V}}$.

6.6.3 Une distribution de probabilités déviée par l'intégration de connaissances

Pour comprendre les résultats, nous observons des exemples avec une attention particulière aux distributions de probabilités des modèles. Nous cherchons à expliquer pourquoi un modèle attribue une probabilité plus élevée à un autre sens et si l'ajout d'informations influence ces résultats.

Noms. Nous prenons en exemple, l'instance suivante qui a été correctement prédite par tous les modèles `EWISER` en évaluation restreinte :

« Principal **avocat** de la mise en œuvre des nouveaux instruments, la France soutient la Commission dans sa volonté de pérenniser l'effet structurant des projets intégrés, des réseaux d'excellence et des actions de coordination ERANET, dont la continuité au sein du 7e PCRD doit être assurée. »

Le lexème **avocat** est annoté avec l'acception `avocat 1.II` du vocable `avocat 1`[62]. Le vocable `avocat 1` présente deux acceptions dans le `RL-fr` :

— `avocat 1.I` : individu qui pratique un métier,

— `avocat 1.II` : individu qui a un certain comportement.

Le tableau 6.15 présente la probabilité du *gold*, notée $P(\hat{s}_{\mathcal{S}(w)})$ et du sens avec la probabilité maximale, $P(\hat{s}_{\mathcal{V}})$ ainsi que la distance en nombre de pas entre $\hat{s}_{\mathcal{S}(w)}$ et $\hat{s}_{\mathcal{V}}$ dans le `RL-fr`.

Dans un premier temps, nous remarquons qu'`EWISER-A` prédit le *gold* avec la probabilité maximale et que les probabilités du *gold* sont plus faibles dans les modèles `EWISER+A`, voire quasi nulles pour `EWISER+A STRUCT`, `EWISER+A SEM` et `EWISER+A SEM*`, qui présentent des probabilités inférieures à 10^{-4}. Ce résultat remet en question la validité d'une prédiction correcte avec une probabilité aussi faible, dans la mesure où la différence de probabilité est minime entre les différents sens possibles du lexème-cible.

Dans un second temps, nous observons le sens prédit avec la probabilité maximale, $\hat{s}_{\mathcal{V}}$ et sa distance dans le `RL-fr` avec le sens *gold*. La distance est calculée uniquement à partir des relations encodées par des fonctions lexicales et non les relations de copolysémie.

Notre hypothèse principale est que la matrice d'adjacence influence la distribution de probabilités vers des sens autres que les sens possibles du lexème-cible, et potentiellement des sens proches dans le graphe. On observe toutefois sur l'exemple que les sens prédits avec la probabilité maximale paraissent éloignés sémantiquement du *gold*, avec une distance moyenne entre $\hat{s}_{\mathcal{V}}$ et $\hat{s}_{\mathcal{S}(w)}$ de 4 pas.

Nous calculons la distance moyenne entre le sens *gold* et le sens $\hat{s}_{\mathcal{V}}$ pour chaque modèle sur toutes les instances où $gold \neq \hat{s}_{\mathcal{V}}$, afin de vérifier si l'ajout de la matrice permet au modèle de prédire des sens plus proches du *gold*. Les distances moyennes vont de 2,56 à 3,45 pour les

62. `https://spiderlex.atilf.fr/fr/q/*avocat**1*`

Modèle	$P(\hat{s}_{\mathcal{S}(w)})$	$\hat{s}_{\mathcal{V}}$	$P(\hat{s}_{\mathcal{V}})$	Dist($\hat{s}_{\mathcal{S}(w)}$, $\hat{s}_{\mathcal{V}}$)
`EWISER-A`	0,2580	`avocat 1.II`	0,2580	—
`EWISER+A RANDOM`	0,0282	`garante`	0,2014	4
`EWISER+A RANDOM*`	0,0545	`garante`	0,1717	4
`EWISER+A STRUCT`	$\leq 10^{-4}$	`citadelle II`	0,4995	4
`EWISER+A STRUCT*`	0,0004	`train 1.V.1`	0,4615	5
`EWISER+A SEM`	$\leq 10^{-4}$	`citadelle II`	0,9950	4
`EWISER+A SEM*`	$\leq 10^{-4}$	`citadelle II`	0,8799	4
`EWISER+A COS`	0,0116	`garante`	0,2067	4
`EWISER+A COS*`	0,0391	`garante`	0,5343	4

TABLE 6.15 – Prédictions et probabilités des modèles `EWISER` sur l'instance cit6208 ayant pour *gold* = `avocat 1.II`. Dist($\hat{s}_{\mathcal{S}(w)}$, $\hat{s}_{\mathcal{V}}$) représente la distance entre $\hat{s}_{\mathcal{S}(w)}$ et $\hat{s}_{\mathcal{V}}$ dans le `RL-fr` en nombre de pas.

modèles `EWISER+A`, contre une distance moyenne de 2,38 pour `EWISER-A`, ce qui ne nous permet pas de valider l'hypothèse pour les noms.

Verbes. Concernant les verbes, nous avons noté que l'ajout de plongements de graphe *via* la matrice A permet de meilleures performances par rapport à `EWISER-A`, avec 25 instances correctement prédites en plus avec `EWISER+A COS`.

Nous cherchons à observer l'influence de la matrice sur les distributions de probabilités. Le tableau 6.16 présente des statistiques sur les instances correctement prédites avec la probabilité maximale *gold* = $\hat{s}_{\mathcal{V}}$, pour `EWISER-A` et `EWISER+A COS`.

À première vue, les résultats ne montrent pas qu'`EWISER+A COS` prédit le sens correct avec des probabilités plus élevées, les probabilités moyennes étant autour de 0,73 pour les deux modèles. Nous remarquons également que les deux modèles peuvent prédire le sens correct avec des probabilités minimale et maximale très extrêmes.

	`EWISER-A`	`EWISER+A COS`
Nombre d'instances dans le test	873	
Nombre d'instances *gold* = $\hat{s}_{\mathcal{V}}$	332 (38,03 %)	342 (39,29 %)
Probabilité moyenne du *gold*	0,7279	0,7383
Probabilité maximale du *gold*	0,9996	0,9991
Probabilité minimale du *gold*	0,0754	0,1022

TABLE 6.16 – Statistiques sur les instances prédites correctement avec la probabilité maximale par `EWISER-A` et `EWISER+A COS`.

Nous prenons une instance en exemple afin d'observer les distributions de probabilités des deux modèles. Les figures 6.8 et 6.9 présentent les 10 premiers sens classés selon leurs probabilités pour cette instance, respectivement par `EWISER-A` et `EWISER+A COS`.

Sur cette instance dont le *gold* est `courir V.1`, `EWISER-A` prédit le sens correct avec une probabilité maximale de 0,6026. On remarque que les probabilités sont distribuées sur d'autres sens possibles du lexème.

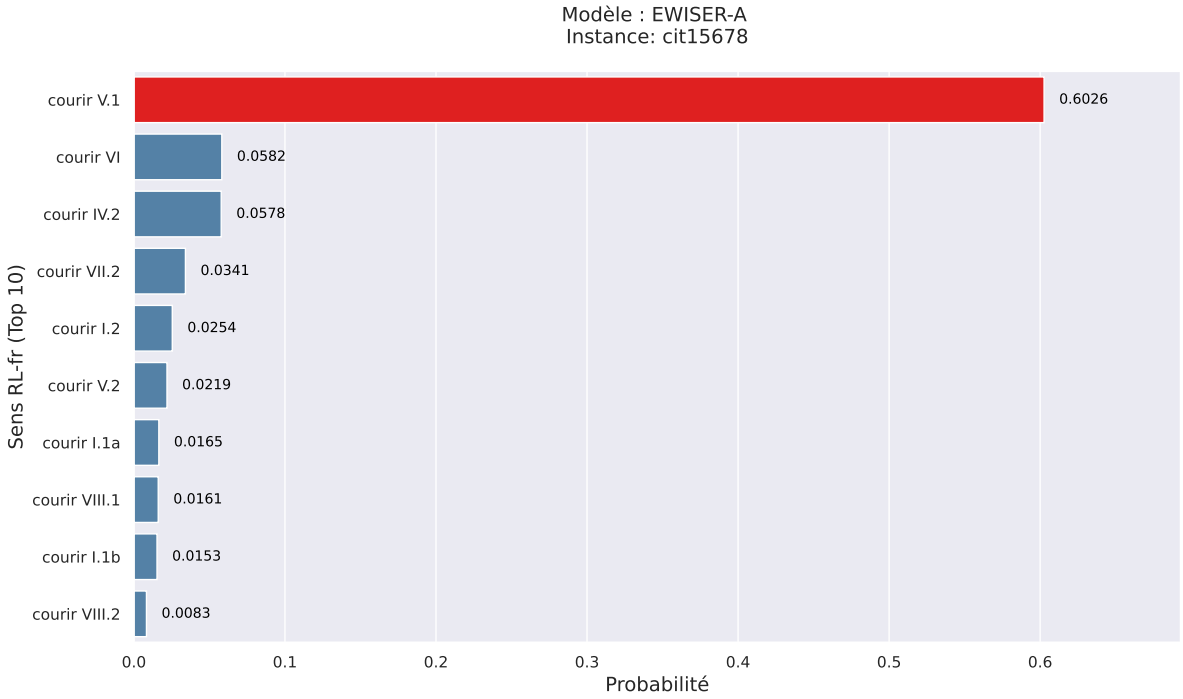


FIGURE 6.8 – Top 10 des sens prédits par `EWISER-A` sur l'instance cit15678. Le sens *gold* est indiqué en rouge.

Avec `EWISER+A COS`, le *gold* est prédit avec une probabilité maximale plus élevée à 0,8803. Une différence importante avec `EWISER-A` est la distribution de probabilités sur des sens autres que les sens possibles du lexème. En effet, `avancer II.2`, `marcher IV` et `filer III` apparaissent dans le top 10.

Il est intéressant de noter qu'`avancer II.2` et `marcher IV` sont liés à `courir V.1` avec une relation de synonymie, ce qui signifierait que la matrice peut effectivement orienter le modèle à distribuer les probabilités sur des sens proches sémantiquement du *gold* sur certaines instances.

6.6.4 Des performances variables selon le niveau de polysémie

Dans cette section, nous étudions l'influence du niveau de polysémie sur les performances des modèles.

Nous supposons que la différence de performances entre les noms et les verbes peut être due au niveau de polysémie qui est généralement plus élevée sur les verbes : 1,61 pour les noms et 2,12 pour les verbes. Sur des données en anglais, Raganato et al. (2017) observent également des performances plus faibles entre les verbes et les noms, attribuant ce résultat au niveau de granularité des verbes qui présentent davantage de sens que les noms communs.

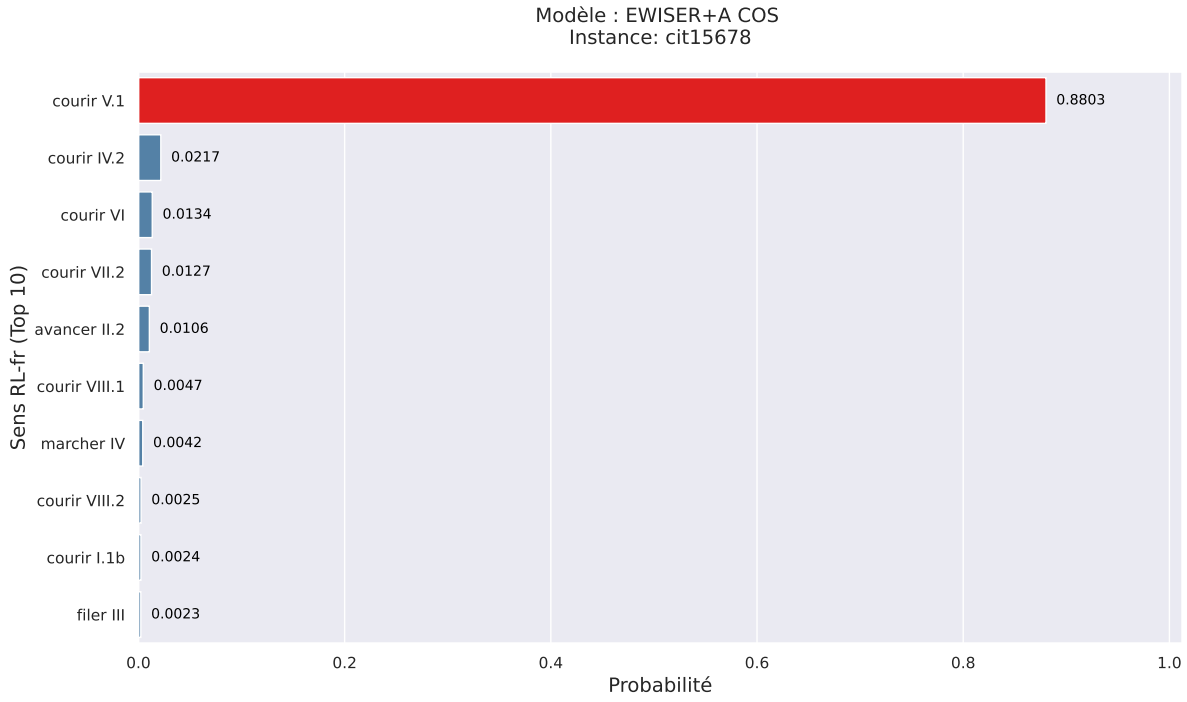


FIGURE 6.9 – Top 10 des sens prédits par `EWISER+A COS` sur l'instance cit15678. Le sens *gold* est indiqué en rouge.

Nous présentons les performances des modèles `EWISER` selon le niveau de polysémie de l'instance sur les noms en figure 6.10 et sur les verbes en figure 6.11.

Nous remarquons que les noms présentent des sens avec une forte polysémie, allant jusqu'à 23 sens possibles (lexème `tête`), ce qui ne nous permet pas de conclure que les modèles atteignent des performances moins élevées en raison du niveau de polysémie.

Contre-intuitivement, la performance des modèles ne diminue pas avec l'augmentation du niveau de polysémie. Les performances sont relativement constantes autour de 0,6 pour les noms communs puis diminuent légèrement sur les sens de niveau de polysémie 21 et 23. Pour les verbes, les valeurs sont en moyenne autour de 0,4, puis présentent des performances variables en fonction du niveau de polysémie.

Nous observons également que les instances pour chaque niveau de polysémie peuvent correspondre à un seul même lexème. Par exemple, les 11 instances présentant un sens avec 23 sens possibles sur les noms correspondent au lexème `tête`. Nous pouvons ainsi supposer que le modèle n'est pas performant pour désambiguïser le lexème en particulier lorsque nous observons un pic de performance élevé ou faible.

Par ailleurs, nous notons que les tendances des performances sont similaires à travers les modèles `EWISER`, avec ou sans matrice d'adjacence. Ainsi, il est difficile de conclure que l'ajout

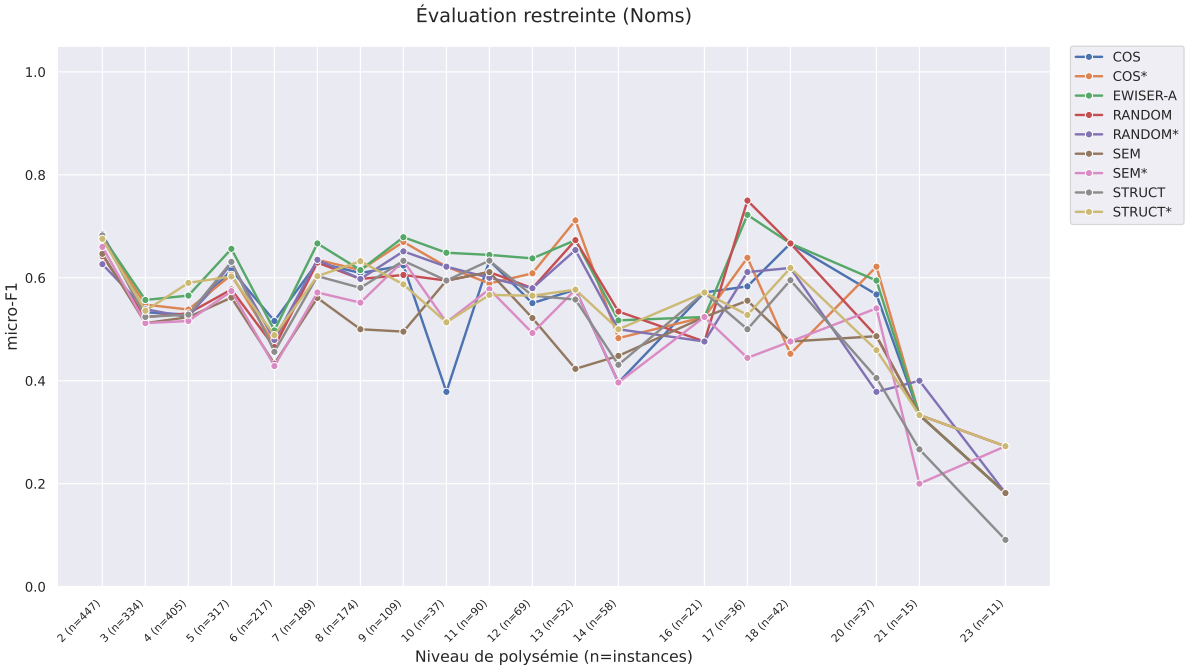


FIGURE 6.10 – Performances des modèles en micro-F1 en évaluation restreinte sur les noms selon le niveau de polysémie. Le nombre d'instances est indiqué entre parenthèses.

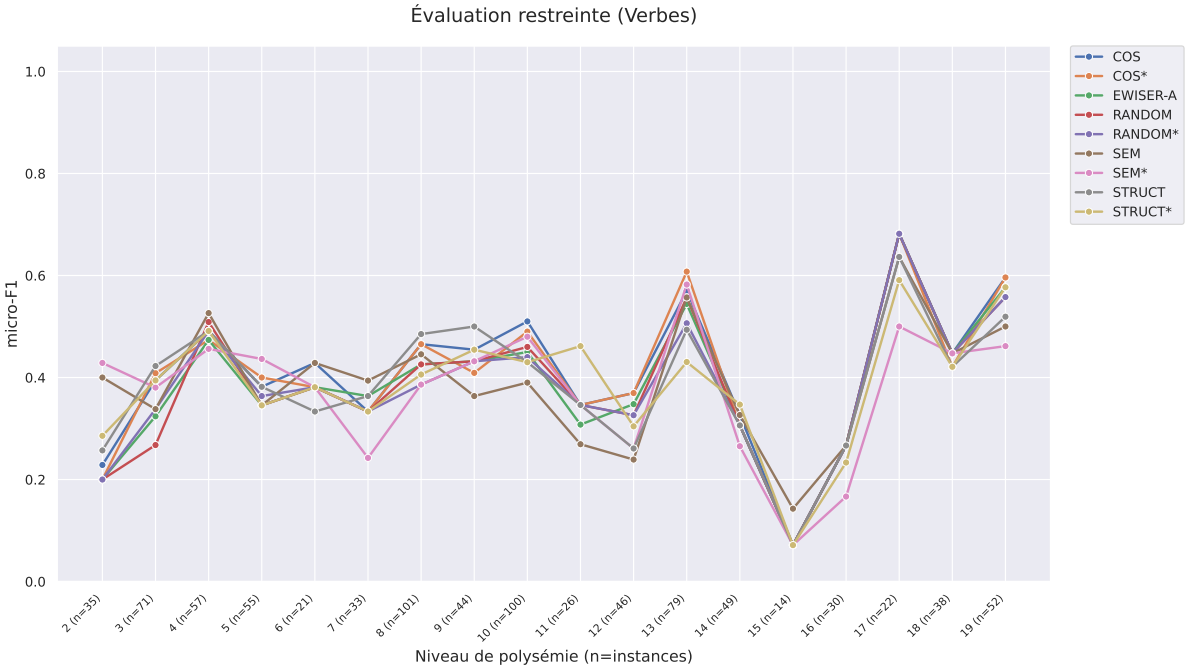


FIGURE 6.11 – Performances des modèles en micro-F1 en évaluation restreinte sur les verbes selon le niveau de polysémie. Le nombre d'instances est indiqué entre parenthèses.

de la matrice a une influence sur la désambiguïsation des lexèmes selon leur niveau de polysémie.

6.6.5 Conclusion de la QR 2

Dans cette question de recherche, nous nous sommes concentrés sur l'influence de l'ajout de connaissances relatives au graphe à travers la matrice d'adjacence A. Nous avons observé que les performances sont moindres avec `EWISER+A` sur les noms mais meilleures sur les verbes avec `EWISER+A COS`.

Tout d'abord, l'ajout de la matrice, quelle que soit la stratégie adoptée, a une influence sur les performances des modèles car cela influe sur les distributions de probabilité. En effet, nous remarquons que les probabilités sont davantage distribuées sur des sens qui ne sont pas des sens possibles des mots-cibles. Ce phénomène est davantage visible avec une évaluation globale où nous pouvons remarquer un écart plus élevé entre la micro-F1 globale et la micro-F1 restreinte dans les modèles `EWISER+A`.

Ensuite, nous avons supposé que les sens qui ne font pas partie des sens possibles des mots-cibles mais qui présentent des probabilités élevées, sont potentiellement des sens proches dans le graphe. Pour les noms, la distance moyenne entre les sens à la probabilité maximale et le *gold* est plus faible dans `EWISER-A` qu'avec les modèles `EWISER+A`, ce qui invalide notre hypothèse. Il semblerait que la matrice A génère du bruit plutôt que des informations utiles au modèle dans le cas des noms.

Pour les verbes, le modèle `EWISER+A COS` présente de meilleures performances avec davantage d'instances correctement prédites. Si les probabilités moyennes des *gold* sont proches avec `EWISER-A` et `EWISER+A COS`, on observe que sur certaines instances, la distribution de probabilités est différente dans la mesure où `EWISER+A COS` peut attribuer des probabilités plus élevées à des sens autre que des sens possibles du lexème-cible, mais qui lui sont sémantiquement proches.

Enfin, les performances des modèles selon le niveau de polysémie ne nous permettent pas de conclure que les performances sont meilleures sur les noms que sur les verbes en raison d'un niveau de polysémie plus faible, certaines instances de noms étant très fortement polysémique et toutefois mieux prédites que les verbes.

6.7 QR 3 : prédiction des sens non vus à l'entraînement

Notre troisième question de recherche se concentre sur les performances des systèmes sur les sens non vus durant l'entraînement. En effet, les modèles ont tendance à prédire des sens vus à l'entraînement, et prédire des sens inconnus constitue une tâche plus difficile car cela nécessite des capacités de généralisation et d'inférence en *zero-shot* (Maru et al., 2022).

Dans cette section, nous cherchons à répondre à deux questions : existe-il une corrélation entre la fréquence d'apparition d'un sens à l'entraînement et la performance du modèle à l'inférence sur les données du `BEL-RL-fr` ? Et si oui, l'intégration de connaissances *via* la matrice A peut-elle améliorer les performances sur les sens non vus à l'entraînement ?

6.7.1 Corrélation entre la distribution des sens à l'entraînement et la performance des modèles

Dans la section 5.6.3, nous avons discuté du biais des modèles à prédire davantage des sens vus à l'entraînement. Nous vérifions dans un premier temps si ce biais est présent dans nos expériences. Pour ce faire, nous calculons la fréquence d'apparition d'un sens dans les données d'entraînement et la performance moyenne en micro-F1 pour chaque sens à l'inférence. Nous utilisons les performances des modèles `EWISER-A` sans la matrice d'adjacence pour les noms et pour les verbes, afin d'observer uniquement l'influence des données d'entraînement et non de l'ajout de la matrice.

Nous calculons une corrélation de Spearman qui permet de mesurer des relations monotones sans supposer de linéarité, contrairement au coefficient de Pearson (Spearman, 2010). Le coefficient de Spearman est compris entre -1 et 1. Une valeur proche de 1 ou -1 indique une corrélation positive ou négative. Une valeur proche de 0 indique qu'il n'existe pas de corrélation entre les variables.

Pour les noms, le coefficient de Spearman est de 0,6331 avec une valeur-p de $1,93 \times 10^{-186}$. Pour les verbes, le coefficient est de 0,6631 avec une p-valeur de $3,43 \times 10^{-70}$. Une corrélation est considérée comme statistiquement significative lorsque la valeur-p est inférieure à 0,05.

Les coefficients de Spearman indiquent qu'il existe une corrélation entre la fréquence d'apparition d'un sens dans l'entraînement et la micro-F1 moyenne sur ce sens à l'inférence : plus un sens est vu à l'entraînement, plus il est correctement prédit par le modèle.

6.7.2 Amélioration des performances des sens non vus avec `EWISER+A`

Nous nous intéressons maintenant à l'influence de l'ajout de connaissances *via* la matrice A sur les sens non vus à l'entraînement. Nous analyserons les performances des modèles `EWISER-A` et `EWISER+A`.

Le tableau 6.17 présente le nombre d'instances dans l'échantillon de test dont le sens *gold* est un sens vu ou non vu à l'entraînement. La proportion d'instances dont le sens correct n'a pas été vu à l'entraînement est de 16 % pour les noms et 25 % pour les verbes.

	Noms	**Verbes**
Nombre total instances	2 660	873
Nombre d'instances avec sens *gold* vu	2 228	626
Nombre d'instances avec sens *gold* non vu	432	217

TABLE 6.17 – Statistiques sur les instances de test dont le sens *gold* est un sens vu ou non vu à l'entraînement.

Les tableaux 6.18 et 6.19 présentent les performances en micro-F1 des modèles `EWISER` en séparant les instances dont le sens *gold* est connu et inconnu de l'entraînement, respectivement sur les noms et les verbes. Nous présentons les résultats avec une évaluation globale et restreinte.

Tout d'abord, nous remarquons qu'en évaluation globale, sur les noms comme sur les verbes, les performances sont très faibles et même nulles pour la majorité des modèles. Cela signifie que les modèles ne sont pas capables de prédire un sens correct inconnu de l'entraînement avec une

probabilité maximale.

Ensuite, nous observons que les performances augmentent en évaluation restreinte sur le sens non vus à l'entraînement, notamment avec les modèles `EWISER+A`.

Plus spécifiquement, `EWISER+A STRUCT` présente une performance de 8,56 % sur les noms (37 instances). En comparaison, `EWISER-A` atteint 0,23 %, ce qui correspond à une seule instance bien prédite. Pour les verbes, `EWISER+A SEM*` prédit correctement 22 instances (10,14 %) tandis que le modèle sans matrice A, `EWISER-A` présente une performance nulle. Les performances restent très faibles, inférieure à 10 % sur tous les modèles.

Par ailleurs, les modèles `EWISER+A` ont des performances plus faibles qu'`EWISER-A` sur les sens vus à l'entraînement sur les noms, ce qui explique pourquoi la performance finale est plus élevée avec `EWISER-A`. Sur les verbes, nous observons toutefois un gain avec `EWISER+A COS` et `EWISER+A COS*` sur les sens vus et non vus à l'entraînement, bien que le nombre de sens non vus bien prédits est plus faible par rapport à `EWISER+A SEM*`.

	Évaluation globale			**Évaluation restreinte**		
Modèle	**Total**	**Vus**	**Non vus**	**Total**	**Vus**	**Non vus**
`EWISER-A`	59,21	70,69	0	61,17	72,98	0,23
`EWISER+A RANDOM`	54,29	64,81	0	57,29	68,36	0,23
`EWISER+A RANDOM*`	51,92	61,98	0	57,52	68,00	3,47
`EWISER+A STRUCT`	44,02	52,51	0,23	57,29	66,74	**8,56**
`EWISER+A STRUCT*`	51,32	61,22	0,23	58,46	68,85	4,86
`EWISER+A SEM`	40,23	47,85	0,93	53,68	62,75	6,94
`EWISER+A SEM*`	41,92	50,04	0	54,66	63,96	6,71
`EWISER+A COS`	52,86	63,11	0	58,23	68,76	3,94
`EWISER+A COS*`	53,53	63,91	0	58,76	69,84	1,62

TABLE 6.18 – Performances (%) en micro-F1 des modèles `EWISER` sur les instances des noms dont le sens correct est vu ou non vu à l'entraînement.

6.7.3 Conclusion de la QR 3

Dans cette troisième question de recherche, nous avons montré que plus un sens est fréquent dans les données d'entraînement, plus les modèles prédisent correctement ce sens à l'inférence. Ce résultat confirme que la prédiction en *zero-shot* sur les sens non vus est donc une tâche plus difficile.

Par ailleurs, nous avons observé l'influence de l'intégration de connaissances dans les modèles `EWISER+A` sur ces sens non vus.

Pour les noms comme pour les verbes, nous notons une augmentation des performances en évaluation restreinte sur les sens non vus à l'entraînement, notamment avec `EWISER+A STRUCT` et `EWISER+A SEM*`.

Pour les noms, nous observons que les sens vus à l'entraînement présentent des performances plus faibles, résultat discuté en section 6.6. Pour les verbes, seuls `EWISER+A COS` et

	Évaluation globale			Évaluation restreinte		
Modèle	**Total**	**Vus**	**Non vus**	**Total**	**Vus**	**Non vus**
EWISER-A	38,03	50,61	0	40,78	54,27	0
EWISER+A RANDOM	37,91	50,46	0	40,55	53,96	0
EWISER+A RANDOM*	36,99	49,24	0	39,97	53,05	0,46
EWISER+A STRUCT	30,93	41,16	0	41,47	52,13	9,22
EWISER+A STRUCT*	33,22	44,05	0,46	40,32	51,37	6,91
EWISER+A SEM	30,81	40,85	0,46	40,55	51,22	8,29
EWISER+A SEM*	31,04	41,31	0	40,32	50,30	**10,14**
EWISER+A COS	39,29	52,29	0	43,76	56,71	4,61
EWISER+A COS*	39,63	52,74	0	43,30	56,86	2,30

TABLE 6.19 – Performances (%) en micro-F1 des modèles EWISER sur les instances des verbes dont le sens correct est vu ou non vu de l'entraînement.

EWISER+A COS* présentent de meilleures performances sur les sens vus et non vus à l'entraînement, donnant ainsi une performance globale plus élevée qu'EWISER-A.

Il semblerait que la matrice d'adjacence A peut orienter les modèles pour les sens non vus à l'entraînement mais peut également générer du bruit pour les sens vus à l'entraînement, en particulier pour les noms.

6.8 QR 4 : analyse des performances des grands modèles de langue

Dans cette quatrième question de recherche, nous nous intéressons plus particulièrement aux résultats obtenus avec les grands modèles de langue en *zero-shot*. Nous cherchons à observer l'influence des différentes stratégies de *prompting* sur les performances des modèles, données en micro-F1 et macro-F1.

Noms. Pour les noms, nous avons testé huit stratégies de *prompting* différentes variant :

- la langue du *prompt* : anglais (EN) ou français (FR),
- la manière de décrire les sens : avec le voisinage lexical dans le RL-fr (lexical) ou un exemple d'utilisation du BEL-RL-fr (usage),
- la génération ou non d'une définition (definition).

Les tableaux 6.20 et 6.21 présentent les performances des quatre modèles en micro-F1 et macro-F1 sur les noms.

Le modèle présentant la micro-F1 la plus élevée est Mistral-7b avec 33,92 %, avec un *prompt* en français décrivant les sens avec un cas d'utilisation dans le BEL-RL-fr. Avec une macro-F1, le modèle le plus performant est Qwen2.5-7b (32,98 %) avec un *prompt* en français décrivant les sens avec le voisinage lexical dans le RL-fr.

Nous remarquons que générer une définition intermédiaire ne permet pas d'obtenir de meilleurs résultats quelle que soit la stratégie de description des sens.

Stratégie	**Lexical**				**Lexical + Definition**			
Langue	**EN**		**FR**		**EN**		**FR**	
	M-F1	**m-F1**	**M-F1**	**m-F1**	**M-F1**	**m-F1**	**M-F1**	**m-F1**
`Qwen2.5-7b`	30,82	29,70	**32,98**	24,14	30,76	29,25	31,30	30,94
`Gemma3-4b`	24,40	28,91	26,64	31,05	26,59	31,58	28,11	32,97
`Llama3.1-8b`	27,78	31,88	26,00	29,32	30,96	29,40	28,00	33,23
`Mistral-7b`	32,93	29,59	30,29	21,32	29,11	31,02	28,42	29,10

TABLE 6.20 – Performances (%) en macro-F1 (M-F1) et micro-F1 (m-F1) des grands modèles de langue sur les noms avec les stratégies **lexical** et **lexical_def**.

Stratégie	**Usage**				**Usage + Definition**			
Langue	**EN**		**FR**		**EN**		**FR**	
	M-F1	**m-F1**	**M-F1**	**m-F1**	**M-F1**	**m-F1**	**M-F1**	**m-F1**
`Qwen2.5-7b`	30,54	25,18	28,47	20,06	31,04	27,59	30,65	26,69
`Gemma3-4b`	27,33	28,02	27,43	28,11	29,13	29,23	31,16	31,86
`Llama3.1-8b`	28,76	24,92	31,78	24,80	31,85	23,76	32,73	28,37
`Mistral-7b`	32,64	33,66	32,59	**33,92**	33,50	31,94	33,22	27,68

TABLE 6.21 – Performances (%) en macro-F1 (M-F1) et micro-F1 (m-F1) des grands modèles de langue sur les noms selon les stratégies **usage** et **usage_def**.

Verbes. Concernant les verbes, nous avons testé seulement quatre stratégies de *prompting* (cf. Tableau 6.22). La stratégie **usage** n'a pas pu être appliquée, les exemples d'utilisation dans le `BEL-RL-fr` ne couvrant pas tous les sens possibles des verbes-cibles.

Comme pour les noms, nous remarquons que le modèle le plus performant dépend de la métrique considérée : `Qwen2.5-7b` présente la macro-F1 la plus élevée avec un prompt en anglais suivant la stratégie lexicale avec génération de définition (**en_lexical_definition**) et `Gemma3-4b` présente la micro-F1 la plus élevée avec la même stratégie mais en français (**fr_lexical_def**).

Les résultats sont légèrement plus élevés avec la stratégie **lexical_def**, indiquant que la génération d'une définition pourrait permettre d'améliorer les prédictions du modèle.

Stratégie	**Lexical**				**Lexical + Definition**			
Langue	**EN**		**FR**		**EN**		**FR**	
	M-F1	**m-F1**	**M-F1**	**m-F1**	**M-F1**	**m-F1**	**M-F1**	**m-F1**
`Qwen2.5-7b`	23,63	21,53	23,71	18,21	**26,87**	23,25	23,09	20,73
`Gemma3-4b`	16,65	20,39	18,69	21,99	18,84	22,11	20,62	**23,37**
`Llama3.1-8b`	22,67	17,87	17,02	19,70	23,42	19,82	18,53	18,90
`Mistral-7b`	17,74	19,36	22,99	14,78	18,50	19,24	22,13	17,87

TABLE 6.22 – Performances (%) en macro-F1 (M-F1) et micro-F1 (m-F1) des grands modèles de langue sur les verbes selon les stratégies **lexical** et **lexical_def**.

6.9 Discussion et conclusion

Dans ce chapitre, nous nous sommes intéressés à la désambiguïsation lexicale des noms communs et de verbes issus d'exemples lexicographiques du BEL-RL-fr, un corpus annoté en sens du RL-fr. N'ayant pas été créé initialement pour la tâche, le corpus est atypique dans sa distribution des lexèmes et des sens qui ne suit pas celle d'un texte naturel.

Les expériences présentées se focalisent sur les performances de différents types de modèles de désambiguïsation lexicale, notamment EWISER, auquel nous intégrons des plongements de graphe du RL-fr générés dans des expériences précédentes (cf. Chapitre 4).

Par ailleurs, ce chapitre consacre une partie importante à l'évaluation des performances des données, à travers des métriques de macro-F1 et micro-F1 et une analyse des distributions de probabilités pour les modèles EWISER.

Discussion. Notre étude s'est articulée autour de quatre questions de recherche, pour lesquelles nous fournissons les conclusions suivantes.

QR 1 : Quels systèmes produisent les meilleurs performances sur le corpus BEL-RL-fr dans un cadre d'évaluation classique ?

Dans la section 6.5, nous fournissons les performances des différents systèmes en micro-F1 et macro-F1. Les modèles supervisés neuronaux présentent les meilleures performances, comparés aux *baselines* proposées et aux grands modèles de langue. Plus précisément, EWISER-A présente les meilleures performances sur les noms communs avec une micro-F1 de 61,17 % et EWISER+A COS, sur les verbes avec une micro-F1 de 43,76 %.

Les performances calculées avec une micro-F1 et une macro-F1 présentent un écart d'environ 9 points en raison du biais de la composition des données d'évaluation : une grande majorité des sens peu fréquents (une instance seulement) est incorrectement prédite par les modèles, ce qui fait diminuer davantage la valeur de la macro-F1.

QR 2 : Quelle est l'influence de l'intégration de connaissances relatives au graphe à travers la matrice d'adjacence dans le modèle EWISER ?

La matrice d'adjacence A dans EWISER permet d'encoder des poids entre des sens reliés dans le réseau RL-fr. Différentes stratégies sont adoptées pour faire varier la valeur des poids (cf. Section 6.3.2.2). Dans la section 6.6, nous avons cherché à vérifier si l'ajout de la matrice influence positivement ou négativement les performances du modèle.

Afin d'analyser les performances des modèles, nous proposons un nouveau cadre d'évaluation, appelé *évaluation globale*, qui consiste à vérifier si le *gold* est prédit avec la probabilité maximale sur tout le vocabulaire. Avec une évaluation classique, que nous appelons *évaluation restreinte*, le sens prédit est le sens avec la probabilité maximale parmi les sens possibles du lexème-cible.

Nous observons ainsi que les modèles EWISER présentent des performances plus faibles avec une évaluation globale, indiquant que sur certaines instances, le sens *gold* n'est pas prédit avec la probabilité maximale sur tout le vocabulaire. Avec une évaluation restreinte, le sens prédit avec la probabilité maximale sur les sens possibles peut présenter une probabilité relativement faible lorsque l'on considère les probabilités sur tout le vocabulaire.

Sur les noms, nous observons une diminution des résultats après l'intégration de la matrice A, quelle que soit la stratégie adoptée. L'analyse des distributions de probabilités avec les deux types d'évaluation nous permet de montrer que l'intégration de la matrice modifie la distribution de probabilités vers des sens autres que les sens possibles du lexème-cible. Nous supposons que les modèles `EWISER+A` attribuent des probabilités élevés à des sens proches dans le graphe du sens correct, mais l'hypothèse n'a pas été vérifiée.

Cependant, sur les verbes, l'intégration des poids à partir de la similarité cosinus des plongements de graphe permet une amélioration des performances : le nombre d'instances correctement prédites, ainsi que le nombre d'instances correctement prédites avec la probabilité maximale sont plus élevés.

Enfin, nous avons émis l'hypothèse que les différences de performances entre les noms et les verbes pouvaient être dues à un niveau de polysémie plus élevé pour les verbes. Toutefois, les résultats ne nous permettent pas de confirmer cette hypothèse, dans la mesure où des noms communs très fortement polysémiques sont correctement prédits par les modèles. Nous pouvons supposer que la différence de performances peut provenir de la quantité de données d'entraînement qui est bien plus grande pour les noms communs, avec 13 645 instances contre 5 484 pour les verbes. De plus, la proportion de sens inconnus de l'entraînement est plus élevée pour les verbes (25 %) que les noms communs (16 %), ce qui requiert davantage de prédiction en *zero-shot* aux modèles.

QR 3 : L'intégration de connaissance dans `EWISER` améliore-t-elle les performances sur les sens non vus à l'entraînement ?

Un biais connu des modèles supervisés de désambiguïsation lexicale est la tendance à prédire des sens vus à l'entraînement (Maru et al., 2022).

En Section 6.7, nous avons démontré dans un premier temps, qu'il existait une corrélation entre la distribution des sens du `BEL-RL-fr` dans les données d'entraînement et la performance des modèles : plus un sens apparaît à l'entraînement, mieux il est prédit.

Dans un second temps, en observant les performances des modèles `EWISER` sur les sens vus et non vus à l'entraînement, nous montrons que l'ajout de la matrice permet de légèrement augmenter les résultats sur les sens non vus, bien que les performances restent très faibles. Ce résultat prouve les limites des modèles à prédire en *zéro-shot* et le potentiel d'intégrer des connaissances extérieures pour améliorer les performances sur ces instances.

QR 4 : Quelle est l'influence des stratégies de *prompting* sur les performances des grands modèles de langue génératifs ?

Dans la section 6.8, nous nous sommes concentrés sur les performances des grands modèles de langue génératifs, et plus particulièrement l'influence des stratégies de *prompting*.

Nous avons testé différentes stratégie de *prompting* sur les grands modèles de langue. Les résultats ne sont toutefois pas significatifs en termes de différence selon les *prompts*, les performances ne dépassant pas les 34 % de micro et macro-F1 sur les noms et 27 % sur les verbes. Il serait toutefois intéressant de tester des grands modèles de langue de plus grande taille, qui produisent de meilleures performances sur la désambiguïsation lexicale (Meconi et al., 2025).

Concernant les performances générales des grands modèles de langue, nous avons observé qu'elles sont bien plus faibles que celles des modèles `EWISER`, indiquant des difficultés à désambiguïser les exemples annotés en sens du `RL-fr` en *zero-shot*. Contrairement aux précédentes expériences de désambiguïsation lexicale (Basile et al., 2025; Meconi et al., 2025), les sens ne sont pas décrits par des définitions mais par des listes de lexèmes voisins ou des exemples d'usage, ce qui complexifie la tâche dans un premier temps.

Dans un second temps, le `RL-fr` constitue un inventaire de sens de granularité fine, où la distinction entre deux sens n'est pas triviale, même pour un locuteur français non-expert en lexicographie. Cette tendance a également été observée par Petrov et al. (2025) sur la tâche de résolution d'analogie, où les grands modèles de langue présentent des performances plus faibles sur les données du `RL-fr` que sur un jeu de données présentant une granularité moins fine.

Conclusion. Les expériences présentées dans ce chapitre insistent sur trois volets différents :

— **l'évaluation de l'intégration externe de connaissances** : bien que l'ajout d'une matrice de poids n'ait pas été concluante pour les noms, nous avons cherché à exploiter les plongements de graphe qui permettent d'encoder des informations structurelles sur le graphe. Une limite de notre approche réside dans l'utilisation de ces plongements pour modéliser la proximité entre deux sens à l'aide d'une similarité cosinus entre leurs plongements de graphe respectifs. Une telle opération peut en effet restreindre l'expressivité des plongements de graphe, en les réduisant à une unique valeur de similarité. Il serait alors intéressant de trouver un autre moyen d'intégrer ces connaissances, de la même manière que l'intégration de plongement de sens (Bevilacqua and Navigli, 2020).

— **l'importance de la métrique** : il est difficile, voire impossible, de résumer les performances d'un système avec une seule métrique. Une métrique représente une interprétation agrégée des résultats. Bien qu'une valeur chiffrée de la performance permette de comparer les systèmes entre eux et de juger à vue d'œil celui qui paraît être le plus performant, cette manière d'évaluer peut en réalité masquer des comportements différents, qui ne se révèlent qu'à travers une analyse qualitative des résultats.

— **aller au-delà de la classification** : nous avons cherché à regarder au-delà de la simple tâche d'attribuer le sens correct parmi des sens possibles. L'objectif était d'analyser plus spécifiquement les prédictions du modèle, notamment la distribution de probabilités générée qui est généralement masquée par la métrique finale attestant de la performance du modèle. Une autre perspective de travaux serait d'observer si des phénomènes similaires apparaissent dans les *benchmarks* classiques utilisés pour l'évaluation de la désambiguïsation lexicale.

Conclusion

Conclusion et perspectives

Les ressources lexicales occupent une place centrale en lexicographie et en traitement automatique des langues (TAL), d'une part pour leur description du vocabulaire d'une langue et d'autre part pour leur usage dans des systèmes automatiques.

Cette thèse s'est attachée à explorer les interactions entre les ressources lexicales et le TAL en observant à la fois comment les techniques de TAL peuvent pallier l'incomplétude des ressources lexicales en les enrichissant, et comment ces ressources peuvent en retour améliorer les systèmes de TAL.

1 Enrichissement mutuel entre ressources lexicales et systèmes de TAL

Le premier volet « TAL pour les ressources », est abordé dans la seconde partie de la thèse sur l'enrichissement de réseaux lexico-sémantiques du français à partir de modèles de prédiction de liens. Bien que la problématique de l'incomplétude soit avancée dans la majorité des travaux, nous remarquons que très peu aboutissent à un enrichissement concret du graphe d'origine. Nous avons donc cherché à répondre à notre troisième question de recherche (QR 3) : **un modèle de prédiction de liens peut-il prédire de potentiels nouveaux triplets pour un réseau lexico-sémantique parcimonieux ?**

Dans un premier temps, nous avons introduit cette partie avec le chapitre 3 qui présente un état de l'art de la tâche de complétion de graphe par prédiction de liens, où nous avons observé la tendance de la majorité des travaux à se focaliser sur la conception de modèles de prédiction de liens et leurs performances, plutôt que sur la ressource elle-même.

Dans un second temps, afin d'adopter un point de vue centré sur l'enrichissement de la ressource, nous avons analysé les prédictions de `CompGCN-ConvE`, le modèle le plus performant, pour extraire de potentiels nouveaux triplets. Notre travail, décrit dans le chapitre 4, a abouti à plusieurs contributions (Choi et al., 2024b,a) :

— La comparaison de sept modèles de prédiction de liens sur deux réseaux lexico-sémantiques, `RezoJDM16k` (Mirzapour et al., 2022) et `RL-fr` (Lux-Pogodalla and Polguère, 2011). Pour `RezoJDM16k`, nous avons testé le modèle `CompGCN-ConvE` qui présente les meilleures performances en comparaison aux modèles testés par Mirzapour et al. (2022). Ce modèle produit également les meilleures performances sur le `RL-fr`.

— La génération de prédictions robustes *via* la technique de Monte-Carlo Dropout (Gal and Ghahramani, 2016) permettant d'estimer l'incertitude du modèle le plus performant.

— L'extraction de nouveaux triplets pertinents pour le `RL-fr`, par le biais d'un score de confiance évalué manuellement.

Dans cette étude, nous avons cherché à exploiter toutes les prédictions produites par le modèle de prédiction de liens, sans nous arrêter à l'évaluation de la performance du modèle qui consiste à observer uniquement sa capacité à prédire les triplets corrects avec un score élevé. En estimant la robustesse du modèle à prédire les triplets, nous avons pu proposer aux lexicographes du `RL-fr` près de 400 triplets jugés pertinents avec un score de confiance de 95 % pour un potentiel enrichissement de la ressource.

La troisième partie de la thèse se consacre au volet « ressources pour le TAL » avec la quatrième question de recherche (QR 4) : **l'intégration de plongements de graphe, appris à partir d'un réseau lexico-sémantique, peut-elle améliorer les performances d'un modèle de désambiguïsation lexicale du français ?**

La partie s'ouvre avec le chapitre 5 qui présente un état de l'art de la tâche de désambiguïsation lexicale, tâche historique et largement étudiée par la communauté, à travers diverses campagnes d'évaluation. Dans ce chapitre, nous avons décrit des jeux de données inscrits en tant que données de référence depuis plusieurs années, en particulier pour l'anglais, ainsi que la succession de modèles permettant d'atteindre des performances « proches » de l'accord inter-annotateurs sur la tâche de désambiguïsation lexicale (Navigli, 2009).

Dans le chapitre 6, nous avons évalué différents systèmes de désambiguïsation lexicale sur des données du français peu utilisées pour cette tâche. En effet, le `BEL-RL-fr` (Lux-Pogodalla, 2014) est un corpus d'exemples lexicographiques du français, initialement construit pour illustrer des sens du `RL-fr`. Le corpus étant construit artificiellement, la distribution des sens ne reflète pas celle présente dans des textes naturels. Par ailleurs, `BEL-RL-fr` constitue le seul corpus disponible annoté en sens du `RL-fr`, un inventaire de sens relativement rare dans le cadre de la désambiguïsation lexicale.

Nos expériences s'inscrivent dans la continuité des travaux de Sinha et al. (2022) qui initient l'utilisation du modèle neuronal supervisé `EWISER` (Bevilacqua and Navigli, 2019) pour désambiguïser les noms communs et les verbes dans le `BEL-RL-fr`. Ce modèle permet d'intégrer des informations relatives au graphe en encodant des poids entre les sens, selon les relations qui les lient dans le graphe, à travers une matrice d'adjacence. Notre travail apporte les contributions suivantes :

- L'ajout d'une *baseline* basée sur la comparaison entre le voisinage lexical d'un sens dans le `RL-fr` et le voisinage lexical dans la phrase à désambiguïser.
- L'intégration des plongements de graphe générés dans le chapitre 4 dans le modèle `EWISER`.
- L'évaluation de quatre grands modèles de langue auto-régressifs sur la désambiguïsation des exemples du `BEL-RL-fr`, avec différentes stratégies de *prompting*.
- Un nouveau cadre d'évaluation qui consiste à vérifier si le modèle `EWISER` prédit le sens correct avec la probabilité maximale parmi tous les sens de tous les mots. Les résultats sont également calculés avec une micro et une macro-F1.
- Une analyse qualitative des résultats sur les sens non vus à l'entraînement et sur l'influence de l'intégration de la matrice.

Les résultats montrent que le modèle neuronal supervisé `EWISER` est plus performant sur les noms communs (micro-F1 de 61,17 %) et sur les verbes (micro-F1 de 43,76 %). Plus spécifiquement, les performances sur les verbes sont améliorées après l'intégration des plongements de graphe mais pas sur les noms communs, pour lesquels les résultats sont meilleurs sans l'ajout de

la matrice d'adjacence. Toutefois, nous avons observé que l'ajout de la matrice permet d'obtenir des performances légèrement plus élevées sur les sens non vus à l'entraînement, ce qui pallie la configuration *zero-shot* pour ces sens.

Par ailleurs, les performances des grands modèles de langue auto-régressifs atteignent en moyenne une micro-F1 de 30 %, indépendamment des stratégies de *prompting*. En l'absence de définition pour décrire les sens, la tâche prend une forme plus complexe avec la description des sens par le voisinage lexical dans le `RL-fr` ou par un cas d'usage dans le `BEL-RL-fr`, ce qui peut expliquer les résultats plus faibles.

2 Nécessité d'évaluer au-delà d'une métrique unique

Dans cette thèse, nous avons accordé une place importante à l'évaluation des systèmes, particulièrement sur la tâche de désambiguïsation lexicale à travers la cinquième question de recherche (QR 5) : **le cadre d'évaluation de la tâche de désambiguïsation lexicale reflète-elle les performances réelles des modèles ?**

Dans un premier temps, nous avons traité cette question dans l'état de l'art sur la désambiguïsation lexicale, présenté dans le chapitre 5. Au-delà de présenter un état de l'art sur la tâche, nous avons discuté du cadre d'évaluation de la tâche qui s'est ancré depuis des années dans un schéma fixe, avec des jeux de données de référence de l'anglais massivement utilisés pour l'évaluation (Raganato et al., 2017), et une métrique unique de micro-F1. Si ces conditions d'évaluation semblent faciliter une comparaison des systèmes, elles favorisent surtout une course à la performance qui ne permet pas de refléter ce que les systèmes produisent réellement (Reiter, 2025). En effet, nous avons montré dans notre analyse, que la valeur de performance peut fortement varier selon quatre types de biais :

— le biais des données d'évaluation : la présence d'instances monosémiques et d'instances admettant plusieurs sens corrects possibles augmente les performances des systèmes.

— le biais des données d'entraînement : les systèmes prédisent davantage des sens vus durant l'entraînement, leurs capacités à prédire en *zero-shot* pouvant être limitées.

— le biais de la métrique : une micro-F1 tend à *booster* les performances lorsque les sens les plus fréquents sont bien prédits. Dans ce cas de figure, les scores sont généralement bien plus faibles avec une macro-F1 (Maru et al., 2022).

— le biais du sens le plus fréquent : une problématique connue en désambiguïsation lexicale est le fait que les systèmes tendent à prédire le sens le plus fréquent (Postma et al., 2016).

À la lumière des biais soulevés dans le chapitre 5, nous avons proposé un cadre d'évaluation élargi et une évaluation qualitative dans nos expériences de désambiguïsation lexicale. Dans le chapitre 6, pour éviter les biais liés aux données, nous avons veillé à retirer les instances monosémiques des données d'évaluation et nous avons calculé les performances sur les sens vus et non vus de l'entraînement.

Par ailleurs, les performances ont été données en micro-F1 et en macro-F1 afin d'observer les différences liées à la distribution des sens.

Enfin, nous avons introduit une nouvelle configuration pour l'évaluation du modèle `EWISER`, appelée *évaluation globale*, qui consiste à vérifier si le sens correct est prédit avec la probabilité maximale sur tout le vocabulaire, c'est-à-dire sur tous les sens de tous les lexèmes. Dans une

évaluation classique, que nous appelons *évaluation restreinte*, le sens prédit est le sens avec la probabilité maximale parmi les sens possibles du lexème-cible.

Cette distinction nous a permis d'observer que le modèle ne prédit pas systématiquement le sens correct avec la probabilité maximale sur tout le vocabulaire, indiquant que le modèle attribue une probabilité plus élevée à un autre sens mais qui peut être proche sémantiquement. L'évaluation globale a notamment mis en évidence l'influence de l'ajout de la matrice d'adjacence qui dévie la distribution de probabilités.

3 Changement de statut des ressources lexicales en TAL

En linguistique comme en TAL, les ressources lexicales constituent des sources d'informations sur le vocabulaire d'une langue, mobilisées à des fins aussi bien descriptives qu'applicatives. En TAL plus spécifiquement, nous avons observé un changement majeur dans leur utilisation, passant d'un statut de matériau fondamental aux systèmes à un statut d'outil d'analyse ou d'évaluation. Nous avons cherché à mesurer ce changement en observant la dynamique de l'utilisation des ressources lexicales à travers la deuxième question de recherche (QR 2) : **comment a évolué la présence des ressources lexicales du français sur une période de 25 ans d'articles scientifiques ? Quelles sont les ressources les plus utilisées et pour quelles applications ?**

Dans le chapitre 2, nous avons mené une analyse bibliométrique sur l'évolution d'une trentaine de ressources lexicales du français de 2001 à 2025 (Choi et al., 2023). En constituant un corpus de 1 687 articles issus de la conférence nationale TALN, nous avons observé la création de nombreuses ressources entre les années 2000 à 2015, témoignant des efforts de la communauté pour permettre au français de devenir une langue bien dotée.

La ressource lexicale la plus présente et utilisée dans les articles de TALN est le `Le`***`fff`*** (Clément et al., 2004; Sagot, 2010), un lexique syntaxique et morphologique des formes fléchies du français. Construite à partir de la fusion de ressources existantes, le `Le`***`fff`*** a été créé spécifiquement pour le TAL et présente une large couverture grâce à son processus de construction semi-automatique.

Concernant les applications, nous avons observé que les ressources lexicales ont pu permettre la création de lexiques spécialisés, d'outils tels que des lemmatiseurs ou des étiqueteurs morpho-syntaxiques. Elles ont également été utilisées pour la correction d'erreurs et l'évaluation de systèmes automatiques.

Toutefois, l'utilisation des ressources a diminué progressivement à partir de 2015, avec l'émergence des plongements lexicaux statiques tels que `word2vec` (Mikolov et al., 2013), puis en 2020 avec les grands modèles de langue de type `BERT` (Devlin et al., 2019) basés sur l'architecture transformeur (Vaswani et al., 2017) et enfin, plus récemment, les grands modèles de langue auto-régressifs. Cette tendance s'explique en grande partie par une évolution dans la représentation de l'information, marquée par le passage de représentations symboliques explicites à des représentations vectorielles, lisibles uniquement par la machine.

Cette problématique a été traitée dans le chapitre 1, à travers la première question de recherche (QR 1) : **comment a évolué la représentation de l'information linguistique dans les ressources lexicales, de la linguistique au TAL ?**

Dans ce chapitre, nous avons décrit l'évolution des ressources lexicales sur deux domaines étroitement liés, la linguistique et le TAL, passant dans un premier temps d'un statut d'empreintes de théories linguistiques à celui de matériau fondamental pour les premiers systèmes automatiques de TAL. Nous avons ainsi distingué les ressources lexicales provenant de la linguistique et celles créées spécifiquement pour le TAL. Pour ces dernières, nous avons distingué les approches de construction par fusion ou par traduction de ressources existantes et celles par myriadisation et *via* des jeux ayant un but (Choi, 2022).

Toutefois, l'essor des représentations vectorielles apprises sur de grandes quantités de textes non structurés a changé les pratiques en TAL, avec le développement de modèles de langue basés sur des réseaux de neurones et l'apprentissage profond. Les informations linguistiques sont inférées dans les régularités des données vues durant l'entraînement des modèles. Bien que les ressources lexicales symboliques ne soient plus au premier plan dans l'élaboration de systèmes pour les langues disposant de grandes quantités de textes, elles occupent néanmoins une place importante en tant que données de référence de qualité, pour l'évaluation et pour la construction de systèmes hybrides (Opitz et al., 2025).

4 Perspectives

En termes de perspectives concrètes sur nos expériences, la diversité des configurations envisageables lors de la mise en œuvre de modèles, qu'il s'agisse de prédictions de liens ou de désambiguïsation lexicale, ouvre la porte à de nombreuses pistes d'amélioration des performances, dont nous discutons dans les conclusions de chaque chapitre.

Cependant, en menant cette thèse, nous avons fait le choix de ne pas viser l'obtention des meilleures performances possibles, mais plutôt de nous intéresser aux productions concrètes des modèles, au-delà de leurs scores, afin d'analyser les facteurs influençant les résultats. En ce sens, l'attention a été davantage portée aux données, au cadre d'évaluation et à la représentation linguistique dans les systèmes, plutôt qu'à leur architecture.

Notamment, bien que cette question soit peu développée dans cette thèse, nous nous sommes également intéressés à la représentation intrinsèque des informations linguistiques dans les plongements de graphe entraînés sur une tâche de prédiction de liens. Nos premières observations indiquent vraisemblablement une absence de corrélation entre la distance entre deux nœuds dans le graphe, mesurée en nombre de pas, et la similarité de leurs plongements respectifs. Il serait ainsi pertinent d'approfondir cette analyse, en prenant en compte l'influence des types de relations (par exemple, synonymie ou antonymie).

Nous avons également longuement discuté de l'influence des données sur la variabilité des performances des modèles. Dans nos expériences sur la prédiction de liens, nous avons comparé deux réseaux lexico-sémantiques très différents, `RezoJDM` et `RL-fr`, afin d'éviter de généraliser une performance à un seul jeu de données. De même, il serait intéressant d'étendre les expériences de désambiguïsation lexicale avec l'intégration de plongements de graphe à des jeux de données plus volumineux que le `BEL-RL-fr`, qu'il s'agisse de données en anglais ou de celles en français créées dans le *benchmark* évaluant `FlauBERT` (Le et al., 2020). Une telle démarche suppose toutefois la génération préalable de plongements de graphe pour chacun des sens présents dans l'inventaire considéré.

Par ailleurs, au regard des limites du cadre d'évaluation classique de la tâche de désambiguïsation lexicale, présentées dans les chapitres 5 et 6, il serait pertinent d'examiner si les systèmes supervisés les plus performants sur les données en anglais maintiennent des performances comparables lorsqu'ils sont évalués dans un cadre d'évaluation globale.

Enfin, si les représentations vectorielles permettent d'encoder des informations contextuelles exploitables par les systèmes automatiques, elles soulèvent néanmoins des défis importants en matière d'interprétabilité, de prise en compte des phénomènes linguistiques rares et des inégalités entre les langues.

Bien que cette thèse se soit concentrée sur les ressources lexicales du français et de l'anglais, deux langues largement dotées en données, les pratiques actuelles du TAL posent des enjeux majeurs pour les langues peu dotées. Les approches dominantes, fondées sur l'exploitation de vastes corpus non structurés, s'avèrent en effet peu adaptées à ces langues, pour lesquelles les ressources lexicales constituent des outils essentiels. Ainsi, même si leur rôle peut sembler s'être amoindri, les ressources lexicales demeurent des données de haute qualité pour la description du vocabulaire d'une langue. Dans un contexte où la qualité des données est de plus en plus questionnée, les ressources lexicales représentent un appui précieux pour mettre en évidence les limites des modèles de langue, enrichir les représentations au moyen d'approches hybrides et soutenir le traitement automatique des langues peu dotées.

À l'interface entre ressources lexicales, représentation linguistique et TAL, cette thèse a visé à mettre en évidence le rôle que continuent de jouer les ressources lexicales dans un paysage aujourd'hui dominé par les approches neuronales. Au-delà de leur utilité pour le TAL, ces ressources constituent des données de qualité, souvent élaborées par des linguistes, indispensables à d'autres domaines et d'autant plus précieuses à une époque marquée par le développement de textes générés artificiellement.

Les usages linguistiques évoluent en permanence et se caractérisent par une grande variété, perceptible entre les langues mais également au sein même d'une langue, entre les variations de l'oral et l'écrit et les variations régionales. Dans ce contexte, il apparaît ambitieux de supposer que les modèles de langue puissent encoder, à partir de seules données massives, toute la finesse de ces phénomènes, d'autant plus que les textes artificiels tendent à amplifier les formulations les plus fréquentes au détriment des usages plus rares.

À l'inverse, les ressources lexicales ne se limitent pas à décrire les formes les plus attestées dans les corpus, mais visent à rendre compte de la richesse et de la diversité du vocabulaire d'une langue dans son ensemble. Elles constituent ainsi un levier essentiel pour une représentation linguistique plus fidèle et plus robuste dans les systèmes de TAL.

Bibliographie

Bibliographie

Abeillé, A. and Barrier, N. (2004). Enriching a French Treebank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbonne, Portugal. European Language Resources Association (ELRA).

Abeillé, A., Clément, L., and Toussenel, F. (2003). Building a Treebank for French. In Abeillé, A., editor, *Treebanks*, pages 165 –187. Kluwer, Dordrecht.

Agirre, E., Bengoetxea, K., Gojenola, K., and Nivre, J. (2011). Improving Dependency Parsing with Semantic Classes. In Lin, D., Matsumoto, Y., and Mihalcea, R., editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, pages 699–703, Portland, États-Unis. Association for Computational Linguistics.

Agirre, E., Lopez de Lacalle, O., Fellbaum, C., Hsieh, S.-K., Tesconi, M., Monachini, M., Vossen, P., and Segers, R. (2010). SemEval-2010 Task 17 : All-Words Word Sense Disambiguation on a Specific Domain. In Erk, K. and Strapparava, C., editors, *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 75–80, Uppsala, Suède. Association for Computational Linguistics.

Agirre, E. and Soroa, A. (2009). Personalizing PageRank for Word Sense Disambiguation. In Lascarides, A., Gardent, C., and Nivre, J., editors, *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 33–41, Athènes, Grèce. Association for Computational Linguistics.

Al-Moslmi, T., Gallofré Ocaña, M., L. Opdahl, A., and Veres, C. (2020). Named Entity Extraction for Knowledge Graphs : A Literature Overview. *IEEE Access*, 8 :32862–32881.

Aloui, C., Ramisch, C., Nasr, A., and Barque, L. (2020). SLICE : Supersense-based Lightweight Interpretable Contextual Embeddings. In Scott, D., Bel, N., and Zong, C., editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3357–3370, En distanciel. International Committee on Computational Linguistics.

Alstott, J., Bullmore, E., and Plenz, D. (2014). powerlaw : A Python Package for Analysis of Heavy-Tailed Distributions. *PLOS ONE*, 9(1) :1–11.

Ambridge, B. and Blything, L. (2024). Large language models are better than theoretical linguists at theoretical linguistics. *Theoretical Linguistics*, 50(1-2) :33–48.

Apidianaki, M. and Sagot, B. (2014). Data-driven Synset Induction and Disambiguation for Wordnet Development. *Language Resources and Evaluation*, 48(4) :655–677.

Aprosio, A. P., Giuliano, C., and Lavelli, A. (2013). Extending the Coverage of DBpedia Properties using Distant Supervision over Wikipedia. In Hellmann, S., Filipowska, A., Barrière,

C., Mendes, P. N., and Kontokostas, D., editors, *Proceedings of the NLP & DBpedia workshop co-located with the 12th International Semantic Web Conference (ISWC 2013)*, volume 1064 of *CEUR Workshop Proceedings*, Sydney, Australie. CEUR-WS.org.

Arcan, M., McCrae, J. P., and Buitelaar, P. (2016). Expanding wordnets to new languages with multilingual sense disambiguation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : Technical Papers*, pages 97–108, Osaka, Japon. The COLING 2016 Organizing Committee.

ATILF (2019). Morphalou. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). DBpedia : A Nucleus for a Web of Open Data. In *Proceedings of the 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference*, ISWC'07/ASWC'07, page 722–735, Berlin, Heidelberg. Springer-Verlag.

Baker, C. F. (2014). FrameNet : A Knowledge Base for Natural Language Processing. In Petruck, M. R. L. and de Melo, G., editors, *Proceedings of Frame Semantics in NLP : A Workshop in Honor of Chuck Fillmore (1929-2014)*, pages 1–5, Baltimore, États-Unis. Association for Computational Linguistics.

Baker, C. F. and Fellbaum, C. (2009). WordNet and FrameNet as Complementary Resources for Annotation. In Stede, M., Huang, C.-R., Ide, N., and Meyers, A., editors, *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 125–129, Suntec, Singapour. Association for Computational Linguistics.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet Project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Balloccu, S., Schmidtová, P., Lango, M., and Dusek, O. (2024). Leak, Cheat, Repeat : Data Contamination and Evaluation Malpractices in Closed-Source LLMs. In Graham, Y. and Purver, M., editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 67–93, St. Julian's, Malte. Association for Computational Linguistics.

Balvet, A., Barque, L., Condette, M.-h., Haas, P., Huyghe, R., Marin, R., and Merlo, A. (2011). La ressource Nomage. Confronter les attentes théoriques aux observations du comportement linguistique des nominalisations en corpus. *Traitement Automatique des Langues*, 52(3) :129–152.

Banerjee, S. and Pedersen, T. (2002). An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '02, page 136–145, Berlin, Heidelberg, Allemagne. Springer-Verlag.

Barba, E., Pasini, T., and Navigli, R. (2021a). ESC : Redesigning WSD with Extractive Sense Comprehension. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *Proceedings of the 2021*

Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, pages 4661–4672, En distanciel. Association for Computational Linguistics.

Barba, E., Procopio, L., and Navigli, R. (2021b). ConSeC : Word Sense Disambiguation as Continuous Sense Comprehension. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1492–1503, Online and Punta Cana, République Dominicaine. Association for Computational Linguistics.

Barque, L., Haas, P., Huyghe, R., Tribout, D., Candito, M., Crabbé, B., and Segonne, V. (2020). FrSemCor : Annotating a French Corpus with Supersenses. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5912–5918, Marseille, France. European Language Resources Association.

Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M., and Zampieri, M. (2019). Findings of the 2019 Conference on Machine Translation (WMT19). In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Martins, A., Monz, C., Negri, M., Névéol, A., Neves, M., Post, M., Turchi, M., and Verspoor, K., editors, *Proceedings of the Fourth Conference on Machine Translation (Volume 2 : Shared Task Papers, Day 1)*, pages 1–61, Florence, Italie. Association for Computational Linguistics.

Basile, P., Musacchio, E., and Siciliani, L. (2024). ITA-SENSE - Evaluate LLMs' ability for ITAlian word SENSE disambiguation : A CALAMITA Challenge. In Dell'Orletta, F., Lenci, A., Montemagni, S., and Sprugnoli, R., editors, *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 1217–1221, Pise, Italie. CEUR Workshop Proceedings.

Basile, P., Siciliani, L., Musacchio, E., and Semeraro, G. (2025). Exploring the Word Sense Disambiguation Capabilities of Large Language Models.

Beguš, G., Dąbkowski, M., and Rhodes, R. (2025). Large Linguistic Models : Investigating LLMs' Metalinguistic Abilities. *IEEE Transactions on Artificial Intelligence*, 6(12) :3453–3467. Publisher Copyright : © 2020 IEEE.

Belinkov, Y. and Glass, J. (2019). Analysis Methods in Neural Language Processing : A Survey. *Transactions of the Association for Computational Linguistics*, 7 :49–72.

Bender, E. M. (2019). The #BenderRule : On naming the languages we study and why it matters. *The Gradient*.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the Dangers of Stochastic Parrots : Can Language Models Be Too Big ? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, États-Unis. Association for Computing Machinery.

Bevilacqua, M., Maru, M., and Navigli, R. (2020). Generationary or "How We Went beyond Word Sense Inventories and Learned to Gloss". In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7207–7221, En distanciel. Association for Computational Linguistics.

Bevilacqua, M. and Navigli, R. (2019). Quasi Bidirectional Encoder Representations from Transformers for Word Sense Disambiguation. In Mitkov, R. and Angelova, G., editors, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 122–131, Varna, Bulgarie. INCOMA Ltd.

Bevilacqua, M. and Navigli, R. (2020). Breaking Through the 80% Glass Ceiling : Raising the State of the Art in Word Sense Disambiguation by Incorporating Knowledge Graph Information. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2864, Online. Association for Computational Linguistics.

Bevilacqua, M., Pasini, T., Raganato, A., and Navigli, R. (2021). Recent Trends in Word Sense Disambiguation : A Survey. In Zhou, Z.-H., editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4330–4338. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Bhattacharyya, P. (2010). IndoWordNet. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, La Valette, Malte. European Language Resources Association (ELRA).

Bi, B., Liu, S., Wang, Y., Mei, L., and Cheng, X. (2024). LPNL : Scalable Link Prediction with Large Language Models. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Findings of the Association for Computational Linguistics : ACL 2024*, pages 3615–3625, Bangkok, Thaïlande. Association for Computational Linguistics.

Biemann, C., Shin, S.-I., and Choi, K.-S. (2004). Semiautomatic Extension of CoreNet using a Bootstrapping Mechanism on Corpus-based Co-occurrences. In *COLING 2004 : Proceedings of the 20th International Conference on Computational Linguistics*, pages 1227–1232, Genève, Suisse. COLING.

Bigeard, E. (2017). Construction de lexiques pour l'extraction des mentions de maladies dans les forums de santé. In *Actes des 24ème Conférence sur le Traitement Automatique des Langues Naturelles. 19es REncontres jeunes Chercheurs en Informatique pour le TAL (RECITAL 2017)*, pages 15–27, Orléans, France. ATALA.

Blache, P. and Rauzy, S. (2008). Influence de la qualité de l'étiquetage sur le chunking : une corrélation dépendant de la taille des chunks. In *Actes de la 15ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*, pages 282–291, Avignon, France. ATALA.

Blanche-Benveniste, C. (2013). L'hypothèse de l'approche pronominale. *Travaux neuchâtelois de linguistique*, 58(58) :119–127.

Blevins, T., Joshi, M., and Zettlemoyer, L. (2021). FEWS : Large-Scale, Low-Shot Word Sense Disambiguation with the Dictionary. In Merlo, P., Tiedemann, J., and Tsarfaty, R., editors,

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume, pages 455–465, En distanciel. Association for Computational Linguistics.

Blevins, T., Levy, O., and Zettlemoyer, L. (2018). Deep RNNs Encode Soft Hierarchical Syntax. In Gurevych, I. and Miyao, Y., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, pages 14–19, Melbourne, Australie. Association for Computational Linguistics.

Blevins, T. and Zettlemoyer, L. (2020). Moving Down the Long Tail of Word Sense Disambiguation with Gloss Informed Bi-encoders. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006–1017, En distanciel. Association for Computational Linguistics.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5 :135–146.

Bollacker, K., Cook, R., and Tufts, P. (2007). Freebase : A Shared Database of Structured General Human Knowledge. In *Proceedings of the 22nd National Conference on Artificial Intelligence - Volume 2*, AAAI'07, page 1962–1963, Vancouver, Canada. AAAI Press.

Bond, F. and Paik, K. (2012). A Survey of WordNets and their Licenses. In *In Proceedings of the 6th Global WordNet Conference (GWC 2012)*, pages 64–71, Matsue, Japon.

Bonfante, G., Guillaume, B., and Perrier, G. (2018). *Application de la réécriture de graphes au traitement automatique des langues*, volume 1 of *Série Logique, linguistique et informatique*. ISTE editions.

Bordes, A., Usunier, N., Garcia-Durán, A., Weston, J., and Yakhnenko, O. (2013). Translating Embeddings for Modeling Multi-Relational Data. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 2787–2795, Red Hook, États-Unis. Curran Associates Inc.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language Models are Few-Shot Learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, En distanciel. Curran Associates, Inc.

Burlot, F. and Yvon, F. (2018). Évaluation morphologique pour la traduction automatique : adaptation au français (Morphological Evaluation for Machine Translation : Adaptation to French). In *Actes de la Conférence TALN. Volume 1 - Articles longs, articles courts de TALN*, pages 61–74, Rennes, France. ATALA.

Cai, H., Zheng, V. W., and Chang, K. C.-C. (2018). A Comprehensive Survey of Graph Embedding : Problems, Techniques, and Applications. *IEEE Transactions on Knowledge and Data Engineering*, 30(9) :1616–1637.

Candito, M., Amsili, P., Barque, L., Benamara, F., De Chalendar, G., Djemaa, M., Haas, P., Huyghe, R., Mathieu, Y. Y., Muller, P., Sagot, B., and Vieu, L. (2014). Developing a French FrameNet : Methodology and First results. In *LREC - The 9th edition of the Language Resources and Evaluation Conference*, Reykjavik, Islande.

Candito, M. and Seddah, D. (2012). Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical (The Sequoia Corpus : Syntactic Annotation and Use for a Parser Lexical Domain Adaptation Method) [in French]. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2 : TALN*, pages 321–334, Grenoble, France. ATALA/AFCP.

Cao, D., Béchet, N., and Marteau, P.-F. (2023). Étude comparative des plongements lexicaux pour l'extraction d'entités nommées en français. In Servan, C. and Vilnat, A., editors, *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : travaux de recherche originaux – articles longs*, pages 94–104, Paris, France. ATALA.

Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Jr., E. R. H., and Mitchell, T. M. (2010). Toward an Architecture for Never-Ending Language Learning. In *Proceedings of the Twenty-Fourth Conference on Artificial Intelligence (AAAI 2010)*, Atlanta, États-Unis.

Chen, M., Tian, Y., Chang, K.-W., Skiena, S., and Zaniolo, C. (2018). Co-training Embeddings of Knowledge Graphs and Entity Descriptions for Cross-lingual Entity Alignment. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 3998–4004. International Joint Conferences on Artificial Intelligence Organization.

Chen, Z., Wang, Y., Zhao, B., Cheng, J., Zhao, X., and Duan, Z. (2020). Knowledge Graph Completion : A Review. *IEEE Access*, 8 :192435–192456.

Chiril, P., Benamara Zitoune, F., Moriceau, V., Coulomb-Gully, M., and Kumar, A. (2019). Multilingual and Multitarget Hate Speech Detection in Tweets. In Morin, E., Rosset, S., and Zweigenbaum, P., editors, *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Volume II : Articles courts*, pages 351–360, Toulouse, France. ATALA.

Choi, H.-S. (2022). État de l'art : Liage de ressources lexicales du français. In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 2 : 24e Rencontres Etudiants Chercheurs en Informatique pour le TAL (RECITAL)*, pages 55–68, Avignon, France. ATALA.

Choi, H.-S., Fort, K., Guillaume, B., and Constant, M. (2023). Des ressources lexicales du français et de leur utilisation en TAL : étude des actes de TALN. In Servan, C. and Vilnat, A., editors, *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 2 : travaux de recherche originaux – articles courts*, pages 23–36, Paris, France. ATALA.

Choi, H.-S., Trivedi, P., Constant, M., Fort, K., and Guillaume, B. (2024a). Au-delà de la performance des modèles : la prédiction de liens peut-elle enrichir des graphes lexico-sémantiques du français ? In Balaguer, M., Bendahman, N., Ho-dac, L.-M., Mauclair, J., G Moreno, J., and Pinquier, J., editors, *Actes de la 31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position*, pages 36–49, Toulouse, France. ATALA et AFPC.

Choi, H.-S., Trivedi, P., Constant, M., Fort, K., and Guillaume, B. (2024b). Beyond Model Performance : Can Link Prediction Enrich French Lexical Graphs ? In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2329–2341, Turin, Italie. ELRA and ICCL.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. The MIT Press, 50 edition.

Chomsky, N. (2002). *Syntactic Structures*. A Mouton classic. Mouton de Gruyter.

Ciaccio, C., Miaschi, A., and Dell'Orletta, F. (2025). Evaluating Lexical Proficiency in Neural Language Models. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T., editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 1267–1286, Vienne, Autriche. Association for Computational Linguistics.

Cimiano, P., Lopez, V., Unger, C., Cabrio, E., Ngonga Ngomo, A.-C., and Walter, S. (2013). Multilingual Question Answering over Linked Data (QALD-3) : Lab Overview. In Forner, Pamela, Muller, Henning, Paredes, Roberto, Rosso, Paolo, Stein, and Benno, editors, *Information Access Evaluation. Multilinguality, Multimodality, and Visualization - 4th International Conference of the CLEF Initiative, CLEF 2013, Valencia, Spain, September 23-26, 2013. Proceedings*, volume 8138, pages 321–332, Valence, Espagne. Springer.

Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-Law Distributions in Empirical Data. *SIAM Review*, 51(4) :661–703.

Clément, L., Sagot, B., and Lang, B. (2004). Morphology Based Automatic Acquisition of Large-coverage Lexica. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbonne, Portugal. European Language Resources Association (ELRA).

Collins, M. (1997). Three Generative, Lexicalised Models for Statistical Parsing. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 16–23, Madrid, Espagne. Association for Computational Linguistics.

Constant, M. and Sigogne, A. (2011). MWU-Aware Part-of-Speech Tagging with a CRF Model and Lexical Resources. In *Proceedings of the Workshop on Multiword Expressions : from Parsing and Generation to the Real World*, pages 49–56, Portland, États-Unis. Association for Computational Linguistics.

Constant, M. and Tolone, E. (2010). A generic tool to generate a lexicon for NLP from Lexicon-Grammar tables. In Gioia, M. D., editor, *Actes du 27e Colloque international sur le lexique et la grammaire (L'Aquila, 10-13 septembre 2008). Seconde partie*, volume 1 of *Lingue d'Europa e del Mediterraneo, Grammatica comparata*, pages 79–93. Aracne. ISBN 978-88-548-3166-7.

Cousin, M. (2023). Vers une implémentation de la théorie sens-texte avec les grammaires catégorielles abstraites. In Candito, M., Gerald, T., and Moreno, J. G., editors, *Actes de CORIA-TALN 2023. Actes des 16e Rencontres Jeunes Chercheurs en RI (RJCRI) et 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉ-CITAL)*, pages 72–86, Paris, France. ATALA.

Danlos, L. (2005). ILIMP : Outil pour repérer les occurences du pronom impersonnel il. In Jardino, M., editor, *Actes de la 12ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*, pages 121–130, Dourdan, France. ATALA.

Danlos, L., Constant, M., and Barque, L. (2016). Improvement of VerbNet-like resources by frame typing. In Hajičová, E. and Boguslavsky, I., editors, *Proceedings of the Workshop on Grammar and Lexicon : interactions and interfaces (GramLex)*, pages 61–70, Osaka, Japon. The COLING 2016 Organizing Committee.

Danlos, L., Nakamura, T., and Pradet, Q. (2014). Toward a French VerbeNet (Vers la création d'un Verbnet français) [in French]. In Stéfanini, M.-H. and Le Pesant, D., editors, *TALN-RECITAL 2014 Workshop FondamenTAL 2014 : Ressources lexicales et TAL - vue d'ensemble sur les dictionnaires électroniques de Jean Dubois et Françoise Dubois-Charlier (FondamenTAL 2014 : Lexical Resources and NLP)*, pages 103–108, Marseille, France. ATALA.

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., and Storms, G. (2019). The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior research methods*, 51(3) :987–1006.

Decadt, B., Hoste, V., Daelemans, W., and van den Bosch, A. (2004). GAMBL, genetic algorithm optimization of memory-based WSD. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 108–112, Barcelone, Espagne. Association for Computational Linguistics.

Delli Bovi, C., Camacho-Collados, J., Raganato, A., and Navigli, R. (2017). EuroSense : Automatic Harvesting of Multilingual Sense Annotations from Parallel Text. In Barzilay, R. and Kan, M.-Y., editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, pages 594–600, Vancouver, Canada. Association for Computational Linguistics.

Denis, P. and Sagot, B. (2010). Exploitation d'une ressource lexicale pour la construction d'un étiqueteur morpho-syntaxique état-de-l'art du français. In *Actes de la 17e conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*, pages 21–30, Montréal, Canada. ATALA.

Denis, P. and Sagot, B. (2012). Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging. *Language Resources and Evaluation*, 46(4) :721–736.

Dettmers, T., Minervini, P., Stenetorp, P., and Riedel, S. (2018). Convolutional 2D knowledge graph embeddings. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18, La Nouvelle-Orléans, États-Unis. AAAI Press.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, États-Unis. Association for Computational Linguistics.

Diwersy, S., Falaise, A., Lay, M.-H., and Souvay, G. (2015). Traitements pour l'analyse du français préclassique. In Lecarpentier, J.-M. and Lucas, N., editors, *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, pages 249–255, Caen, France. ATALA.

Doran, C., Egedi, D., Hockey, B. A., Srinivas, B., and Zaidel, M. (1994). XTAG System - A Wide Coverage Grammar for English. In *COLING 1994 Volume 2 : The 15th International Conference on Computational Linguistics*.

Dubois, J. and Dubois-Charlier, F. (1997). *Les verbes français*. Larousse-Bordas, Paris, France.

Ducel, F., Fort, K., Lejeune, G., and Lepage, Y. (2022). Langues par défaut ? Analyse contrastive et diachronique des langues non citées dans les articles de TALN et d'ACL. In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 144–153, Avignon, France. ATALA.

Dupont, Y. (2017). Exploration de traits pour la reconnaissance d'entités nommées du Français par apprentissage automatique (Feature exploration for French Named Entity Recognition with Machine Learning). In *Actes des 24ème Conférence sur le Traitement Automatique des Langues Naturelles. 19es REncontres jeunes Chercheurs en Informatique pour le TAL (RECITAL 2017)*, pages 42–55, Orléans, France. ATALA.

Edmonds, P. and Cotton, S. (2001). SENSEVAL-2 : Overview. In Preiss, J. and Yarowsky, D., editors, *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 1–5, Toulouse, France. Association for Computational Linguistics.

Eger, S., Youssef, P., and Gurevych, I. (2018). Is it Time to Swish ? Comparing Deep Learning Activation Functions Across NLP tasks. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4415–4424, Bruxelles, Belgique. Association for Computational Linguistics.

Eisele, A. and Chen, Y. (2010). MultiUN : A Multilingual Corpus from United Nation Documents. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, La Valette, Malte. European Language Resources Association (ELRA).

Errica, F., Sanvito, D., Siracusano, G., and Bifulco, R. (2025). What Did I Do Wrong ? Quantifying LLMs' Sensitivity and Consistency to Prompt Engineering. In Chiruzzo, L., Ritter, A., and Wang, L., editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics : Human Language Technologies (Volume 1 : Long Papers)*, pages 1543–1558, Albuquerque, États-Unis. Association for Computational Linguistics.

Eshkol-Taravella, I., Barbedette, A., Liu, X., and Soumah, V.-G. (2022). Classification automatique de questions spontanées vs. préparées dans des transcriptions de l'oral (Automatic Classification of Spontaneous vs). In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 305–314, Avignon, France. ATALA.

Étienne, A., Battistelli, D., and Lecorvé, G. (2020). L'expression des émotions dans les textes pour enfants : constitution d'un corpus annoté (Expressing emotions in texts for children : constitution of an annotated corpus). In Benzitoun, C., Braud, C., Huber, L., Langlois, D., Ouni, S., Pogodalla, S., and Schneider, S., editors, *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles*, pages 205–212, Nancy, France. ATALA et AFCP.

Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., and Smith, N. A. (2015). Retrofitting Word Vectors to Semantic Lexicons. In Mihalcea, R., Chai, J., and Sarkar, A., editors, *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 1606–1615, Denver, États-Unis. Association for Computational Linguistics.

Fauconnier, J.-P. (2015). French Word Embeddings.

Fellbaum, C. (1998). *WordNet : An Electronic Lexical Database*. The MIT Press.

Fillmore, C. J. (1982). Frame Semantics. In of Korea, T. L. S., editor, *Linguistics in the Morning Calm*, pages 111–137. Hanshin Publishing Company, Seoul.

Flesch, M. and Burnett, H. (2025). Mesurer les inégalités de genre en ligne avec le genre grammatical : Une étude du subreddit r/france. In *Actes de CORIA-TALN-RJCRI-RECITAL 2025. Actes des 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : articles scientifiques originaux*, pages 527–542, Marseille, France. ATALA.

Fort, K., Adda, G., and Bretonnel Cohen, K. (2011). Amazon Mechanical Turk : Gold Mine or Coal Mine ? *Computational Linguistics*, pages 413–420.

Fort, K. and Guillaume, B. (2007). PrepLex : un lexique des prépositions du français pour l'analyse syntaxique. In Hathout, N. and Muller, P., editors, *Actes de la 14ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*, pages 205–214, Toulouse, France. ATALA.

François, T., Gala, N., Watrin, P., and Fairon, C. (2014). FLELex : a graded lexical resource for French foreign learners. In *International conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Islande.

Gabay, S., Ortiz Suarez, P., Bawden, R., Bartz, A., Gambette, P., and Sagot, B. (2022). Le projet FREEM : ressources, outils et enjeux pour l'étude du français d'Ancien Régime (The F RE EM project : Resources, tools and challenges for the study of Ancien Régime French). In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 154–165, Avignon, France. ATALA.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a Bayesian Approximation : Representing Model Uncertainty in Deep Learning. In Balcan, M. F. and Weinberger, K. Q., editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, États-Unis. PMLR.

Gala, N. (2013). Ressources lexicales mono- et multilingues : une évolution historique au fil des pratiques et des usages. In *Ressources lexicales : contenu, évaluation, utilisation, évaluation.*, volume 30 of *Lingvisticæ Investigationes Supplementa*, pages 1–42. John Benjamins Publishing.

Gala, N., François, T., Bernhard, D., and Fairon, C. (2014). A model to predict lexical complexity and to grade words (Un modèle pour prédire la complexité lexicale et graduer les mots) [in French]. In Blache, P., Béchet, F., and Bigi, B., editors, *Proceedings of TALN 2014 (Volume 1 : Long Papers)*, pages 91–102, Marseille, France. ATALA.

Gala, N., François, T., and Fairon, C. (2013). Towards a French lexicon with difficulty measures : NLP helping to bridge the gap between traditional dictionaries and specialized lexicons. In *eLex - Electronic Lexicography*, Tallinn, Estonie.

Gala, N. and Rey, V. (2008). POLYMOTS : une base de données de constructions dérivationnelles en français à partir de radicaux phonologiques. In Béchet, F. and Bonastre, J.-F., editors, *Actes de la 15ème conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, pages 91–100, Avignon, France. ATALA.

Galárraga, L., Teflioudi, C., Hose, K., and Suchanek, F. (2015). Fast Rule Mining in Ontological Knowledge Bases with AMIE+. *The VLDB Journal*.

Galárraga, L. A., Teflioudi, C., Hose, K., and Suchanek, F. (2013). AMIE : association rule mining under incomplete evidence in ontological knowledge bases. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, page 413–422, New York, États-Unis. Association for Computing Machinery.

Galkin, M., Berrendorf, M., and Hoyt, C. T. (2022). An Open Challenge for Inductive Link Prediction on Knowledge Graphs. In *Proceedings of The WebConf Workshop on Graph Learning Benchmarks 2022 (WWW'22)*, New York, États-Unis. ACM.

Ganitkevitch, J., Van Durme, B., and Callison-Burch, C. (2013). PPDB : The Paraphrase Database. In Vanderwende, L., Daumé III, H., and Kirchhoff, K., editors, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 758–764, Atlanta, États-Unis. Association for Computational Linguistics.

Gardent, C., Guillaume, B., Falk, I., and Perrier, G. (2005). Le lexique-grammaire de M. Gross et le traitement automatique des langues.

Gardner, M., Talukdar, P. P., Kisiel, B., and Mitchell, T. (2013). Improving Learning and Inference in a Large Knowledge-Base using Latent Syntactic Cues. In Yarowsky, D., Baldwin, T., Korhonen, A., Livescu, K., and Bethard, S., editors, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 833–838, Seattle, États-Unis. Association for Computational Linguistics.

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al. (2023). A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1) :1513–1589.

Gosset, C., Boumedyen Billami, M., Lafourcade, M., Bortolaso, C., and Derras, M. (2021). Extraction automatique de relations sémantiques d'hyperonymie et d'hyponymie dans un corpus métier (Automatic extraction of hypernym and hyponym relations in a professional corpus). In *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 162–170, Lille, France. ATALA.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X. E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein,

J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N. P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., and Ma, Z. (2024). The Llama 3 Herd of Models.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning Word Vectors for 157 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japon. European Language Resources Association (ELRA).

Gross, M. (1975). *Méthodes en syntaxe*. Hermann, Paris, France.

Guo, Y., Shang, G., and Clavel, C. (2025). Benchmarking Linguistic Diversity of Large Language Models. *Transactions of the Association for Computational Linguistics*, 13 :1507–1526.

Hadiwinoto, C., Ng, H. T., and Gan, W. C. (2019). Improved Word Sense Disambiguation Using Pre-Trained Contextualized Word Representations. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5297–5306, Hong Kong, Chine. Association for Computational Linguistics.

Hadouche, F. and Lapalme, G. (2010). Une version électronique du LVF comparée avec d'autres ressources lexicales. *Langages*, 179–180(3) :193–220.

Han, K. (2024). *Generating and answering questions across text and knowledge graphs*. Thèse de doctorat, Université de Lorraine.

Han, X., Cao, S., Lv, X., Lin, Y., Liu, Z., Sun, M., and Li, J. (2018). OpenKE : An Open Toolkit for Knowledge Embedding. In Blanco, E. and Lu, W., editors, *Proceedings of the 2018*

Conference on Empirical Methods in Natural Language Processing : System Demonstrations, pages 139–144, Bruxelles, Belgique. Association for Computational Linguistics.

Haspelmath, M. (2023). Defining the word. *WORD*, 69(3) :283–297.

Hathout, N. (2008). Acquistion of the Morphological Structure of the Lexicon Based on Lexical Similarity and Formal Analogy. In Matveeva, I., Biemann, C., Choudhury, M., and Diab, M., editors, *Coling 2008 : Proceedings of the 3rd Textgraphs workshop on Graph-based Algorithms for Natural Language Processing*, pages 1–8, Manchester, Royaume-Uni. Coling 2008 Organizing Committee.

Hathout, N. and Namer, F. (2014). Démonette, a French derivational morpho-semantic network. *Linguistic Issues in Language Technology*, 11(5) :125–168.

Hathout, N., Namer, F., and Dal, G. (2002). An Experimental Constructional Database : The MorTAL Project. In Boucher, P., editor, *Many Morphologies*, pages 178–209. Cascadilla, Somerville, Mass.

Hathout, N. and Sajous, F. (2016). Wiktionnaire's Wikicode GLAWIfied : a Workable French Machine-Readable Dictionary. In Chair), N. C. C., Choukri, K., Declerck, T., Grobelnik, M., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia. European Language Resources Association (ELRA).

Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., and Klakow, D. (2021). A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 2545–2568, En distanciel. Association for Computational Linguistics.

Henestroza Anguiano, E. and Denis, P. (2011). FreDist : Construction automatique d'un thésaurus distributionnel pour le Français (FreDist : Automatic construction of distributional thesauri for French). In Lafourcade, M. and Prince, V., editors, *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, pages 116–121, Montpellier, France. ATALA.

Hewitt, J. and Manning, C. D. (2019). A Structural Probe for Finding Syntax in Word Representations. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, États-Unis. Association for Computational Linguistics.

Hogan, A., Blomqvist, E., Cochez, M., D'amato, C., Melo, G. D., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngomo, A.-C. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., and Zimmermann, A. (2021). Knowledge Graphs. *ACM Comput. Surv.*, 54(4).

Huang, L., Sun, C., Qiu, X., and Huang, X. (2019). GlossBERT : BERT for Word Sense Disambiguation with Gloss Knowledge. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and*

the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3509–3514, Hong Kong, Chine. Association for Computational Linguistics.

Hubert, N. (2024). *Mesure et enrichissement sémantiques des modèles à base d'embeddings pour la prédiction de liens dans les graphes de connaissances*. Thèse de doctorat, Université de Lorraine.

Hubert, N., Monnin, P., and Paulheim, H. (2023). Beyond Transduction : A Survey on Inductive, Few Shot, and Zero Shot Link Prediction in Knowledge Graphs.

Ide, N., Lenci, A., and Calzolari, N. (2003). RDF Instantiation of ISLE/MILE Lexical Entries. In *Proceedings of the ACL 2003 Workshop on Linguistic Annotation : Getting the Model Right*, pages 30–37, Sapporo, Japon. Association for Computational Linguistics.

Jawahar, G., Sagot, B., and Seddah, D. (2019). What Does BERT Learn about the Structure of Language ? In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italie. Association for Computational Linguistics.

Ji, G., He, S., Xu, L., Liu, K., and Zhao, J. (2015). Knowledge Graph Embedding via Dynamic Mapping Matrix. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, pages 687–696, Pékin, Chine. Association for Computational Linguistics.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2023). Mistral 7B.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, online. Association for Computational Linguistics.

Juzek, T. S. and Ward, Z. B. (2025). Why Does ChatGPT "Delve" So Much ? Exploring the Sources of Lexical Overrepresentation in Large Language Models. In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D., and Schockaert, S., editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6397–6411, Abu Dhabi, Émirats Arabes Unis. Association for Computational Linguistics.

Kahane, S. (2003). The Meaning-Text Theory. In *Dependency and Valency, Handbooks of Linguistics and Communication Sciences*. De Gruyter.

Kanayama, H., Zhao, Y., Iwamoto, R., and Ohko, T. (2024). Incorporating Syntax and Lexical Knowledge to Multilingual Sentiment Classification on Large Language Models. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Findings of the Association for Computational Linguistics : ACL 2024*, pages 4810–4817, Bangkok, Thaïlande. Association for Computational Linguistics.

Kang, C. and Choi, J. (2023). Impact of Co-occurrence on Factual Knowledge of Large Language Models. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics : EMNLP 2023*, pages 7721–7735, Singapour, Singapour. Association for Computational Linguistics.

Kilgarriff, A. (2001). English Lexical Sample Task Description. In Preiss, J. and Yarowsky, D., editors, *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 17–20, Toulouse, France. Association for Computational Linguistics.

Kilgarriff, A. and Palmer, M., editors (1998). *Proceedings of the Pilot SENSEVAL*, Sussex, Royaume-Uni. Association for Computational Linguistics.

Kipf, T. N. and Welling, M. (2017). Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations, ICLR 2017, Conference Track Proceedings*, Toulon, France. OpenReview.net.

Kipper-Schuler, K. (2005). *VerbNet : A broad-coverage, comprehensive verb lexicon.* Thèse de doctorat, University of Pennsylvania.

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K., Kocoń, A., Koptyra, B., Mieleszczenko-Kowszewicz, W., Miłkowski, P., Oleksy, M., Piasecki, M., Łukasz Radliński, Wojtasik, K., Woźniak, S., and Kazienko, P. (2023). ChatGPT : Jack of all trades, master of none. *Information Fusion*, 99 :101861.

Kumar, S., Jat, S., Saxena, K., and Talukdar, P. (2019). Zero-shot Word Sense Disambiguation using Sense Definition Embeddings. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5670–5681, Florence, Italie. Association for Computational Linguistics.

Lafourcade, M. and Joubert, A. (2008). JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes. In *Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, Lyon, France.

Lafourcade, M. and Joubert, A. (2013). Bénéfices et limites de l'acquisition lexicale dans l'expérience JeuxDeMots. In *Ressources lexicales : contenu, évaluation, utilisation, évaluation.*, volume 30 of *Lingvisticæ Investigationes Supplementa*, pages 187–216. John Benjamins Publishing.

Lafourcade, M. and Le Brun, N. (2020). JeuxDeMots : Un réseau lexico-sémantique pour le français, issu de jeux et d'inférences. *Revue Lexique*, 27 :47–86.

Lafourcade, M., Le Brun, N., and Zampa, V. (2014). Les couleurs des gens. In *TALN : Traitement Automatique des Langues Naturelles*, Marseille, France.

Lam, K. N., Al Tarouti, F., and Kalita, J. (2014). Automatically constructing Wordnet Synsets. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, pages 106–111, Baltimore, États-Unis. Association for Computational Linguistics.

Laporte, E. (2010). Le lexique-grammaire est-il exploitable pour le traitement des langues ? In Nakamura, T., Laporte, E., Dister, A., and Fairon, C., editors, *Les Tables. La grammaire du français par le menu. Mélanges en hommage à Christian Leclère*, Cahiers du Cental, pages 207–218. Presses universitaires de Louvain.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., and Schwab, D. (2020). FlauBERT : Unsupervised Language Model Pre-training

for French. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Leech, G. (1992). 100 Million Words of English :The British National Corpus (BNC). *Second Language Research*, 28 :1–13.

Lenat, D. B. (1995). CYC : a large-scale investment in knowledge infrastructure. *Commun. ACM*, 38(11) :33–38.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries : how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, SIGDOC '86, page 24–26, New York, États-Unis. Association for Computing Machinery.

Levin, B. (1993). *English Verb Classes and Alternations : A Preliminary Investigation*. University of Chicago Press.

Levine, Y., Lenz, B., Dagan, O., Ram, O., Padnos, D., Sharir, O., Shalev-Shwartz, S., Shashua, A., and Shoham, Y. (2020). SenseBERT : Driving Some Sense into BERT. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4656–4667, En distanciel. Association for Computational Linguistics.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, États-Unis. Curran Associates Inc.

Lin, Y., Liu, Z., Sun, M., Liu, Y., and Zhu, X. (2015). Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 2181–2187. AAAI Press.

Liu, Y., Zhang, K., Huang, Z., Wang, K., Zhang, Y., Liu, Q., and Chen, E. (2023). Enhancing Hierarchical Text Classification through Knowledge Graph Integration. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics : ACL 2023*, pages 5797–5810, Toronto, Canada. Association for Computational Linguistics.

Logan, R., Liu, N. F., Peters, M. E., Gardner, M., and Singh, S. (2019). Barack's Wife Hillary : Using Knowledge Graphs for Fact-Aware Language Modeling. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5962–5971, Florence, Italie. Association for Computational Linguistics.

Lopez, P. (2008–2023). GROBID. `https://github.com/kermitt2/grobid`.

Loureiro, D. and Jorge, A. (2019). Language Modelling Makes Sense : Propagating Representations through WordNet for Full-Coverage Word Sense Disambiguation. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association*

for Computational Linguistics, pages 5682–5691, Florence, Italie. Association for Computational Linguistics.

Lu, H., Yang, H., Huang, H., Zhang, D., Lam, W., and Wei, F. (2024). Chain-of-Dictionary Prompting Elicits Translation in Large Language Models. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 958–976, Miami, États-Unis. Association for Computational Linguistics.

Lü, Y., Huang, J., and Liu, Q. (2007). Improving Statistical Machine Translation Performance by Training Data Selection and Optimization. In Eisner, J., editor, *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 343–350, Prague, République Tchèque. Association for Computational Linguistics.

Luo, P., Wang, X., Shao, W., and Peng, Z. (2019). Towards Understanding Regularization in Batch Normalization. In *International Conference on Learning Representations*, La Nouvelle-Orléans, États-Unis.

Lux-Pogodalla, V. (2014). Integrating lexicographic examples in a lexical network (Intégration relationnelle des exemples lexicographiques dans un réseau lexical) [in French]. In *Proceedings of TALN 2014 (Volume 2 : Short Papers)*, pages 586–591, Marseille, France. ATALA.

Lux-Pogodalla, V. and Polguère, A. (2011). Construction of a French Lexical Network : Methodological Issues. In *First InternationalWorkshop on Lexical Resources, WoLeR 2011*, pages 54–61, Ljubljana, Slovénie.

Mahdisoltani, F., Biega, J., and Suchanek, F. M. (2013). YAGO3 : A Knowledge Base from Multilingual Wikipedias. In *CIDR*, Asilomar, États-Unis.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. (2024). Dissociating language and thought in large language models : a cognitive perspective. *Trends in Cognitive Sciences*, 28(6) :517–540.

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., and Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48) :30046–30054.

Mariani, J., Francopoulo, G., and Paroubek, P. (2019a). The NLP4NLP Corpus (I) : 50 Years of Publication, Collaboration and Citation in Speech and Language Processing. *Frontiers in Research Metrics and Analytics*, 3.

Mariani, J., Francopoulo, G., Paroubek, P., and Vernier, F. (2019b). The NLP4NLP Corpus (II) : 50 Years of Research in Speech and Language Processing. *Frontiers in Research Metrics and Analytics*, 3.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., Villemonte de La Clergerie, É., Seddah, D., and Sagot, B. (2020). CamemBERT : a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, En distanciel.

Maru, M., Conia, S., Bevilacqua, M., and Navigli, R. (2022). Nibbling at the Hard Core of Word Sense Disambiguation. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 4724–4737, Dublin, Irlande. Association for Computational Linguistics.

Maru, M., Scozzafava, F., Martelli, F., and Navigli, R. (2019). SyntagNet : Challenging Supervised Word Sense Disambiguation with Lexical-Semantic Combinations. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3534–3540, Hong Kong, Chine. Association for Computational Linguistics.

McCarthy, D. and Navigli, R. (2007). SemEval-2007 Task 10 : English Lexical Substitution Task. In Agirre, E., Màrquez, L., and Wicentowski, R., editors, *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, Prague, République Tchèque. Association for Computational Linguistics.

McCoy, R. T., Smolensky, P., Linzen, T., Gao, J., and Celikyilmaz, A. (2023). How Much Do Language Models Copy From Their Training Data ? Evaluating Linguistic Novelty in Text Generation Using RAVEN. *Transactions of the Association for Computational Linguistics*, 11 :652–670.

Meconi, D., Stirpe, S., Martelli, F., Lavalle, L., and Navigli, R. (2025). Do Large Language Models Understand Word Senses ? In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V., editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33885–33904, Suzhou, Chine. Association for Computational Linguistics.

Meilicke, C., Chekol, M. W., Ruffinelli, D., and Stuckenschmidt, H. (2019). Anytime bottom-up rule learning for knowledge graph completion. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, IJCAI'19, page 3137–3143. AAAI Press.

Mel'čuk, I. (1996). Lexical functions in lexicography and natural language processing. *Lexical Functions : A Tool for the Description of Lexical Relations in the Lexicon*, pages 37–102.

Mel'čuk, I., Mel'čuk, I., Beck, D., and Polguère, A. (2012). *Semantics : From Meaning to Text*, volume 1 of *Semantics : From Meaning to Text*. John Benjamins Publishing Company.

Mertens, P. (2010). Restrictions de sélection et réalisations syntagmatiques dans DICOVALENCE Conversion vers un format utilisable en TAL. In Langlais, P. and Gagnon, M., editors, *Actes de la 17e conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*, pages 311–320, Montréal, Canada. ATALA.

Mihalcea, R. and Edmonds, P., editors (2004). *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, Barcelone, Espagne. Association for Computational Linguistics.

Mihalcea, R. and Faruque, E. (2004). SenseLearner : Minimally supervised Word Sense Disambiguation for all words in open text. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 155–158, Barcelone, Espagne. Association for Computational Linguistics.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. In Vanderwende, L., Daumé III, H., and Kirchhoff, K., editors, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 746–751, Atlanta, États-Unis. Association for Computational Linguistics.

Miller, G. A. (1995). WordNet : A Lexical Database for English. *Communications of the ACM*, 38(11) :39–41.

Miller, G. A., Chodorow, M., Landes, S., Leacock, C., and Thomas, R. G. (1994). Using a Semantic Concordance for Sense Identification. In *Human Language Technology : Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In Su, K.-Y., Su, J., Wiebe, J., and Li, H., editors, *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapour. Association for Computational Linguistics.

Mirzapour, M., Ragheb, W., Saeedizade, M. J., Cousot, K., Jacquenet, H., Carbon, L., and Lafourcade, M. (2022). Introducing RezoJDM16k : a French KnowledgeGraph DataSet for Link Prediction. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5163–5169, Marseille, France. European Language Resources Association.

Missud, A., Amsili, P., and Villoing, F. (2020). VerNom : une base de paires morphologiques acquise sur très gros corpus (VerNom : a French derivational database acquired on a massive corpus). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles*, pages 305–313, Nancy, France. ATALA et AFCP.

Mitchell, T., Cohen, W., Hruschka, E., Talukdar, P., Betteridge, J., Carlson, A., Dalvi, B., Gardner, M., Kisiel, B., Krishnamurthy, J., Lao, N., Mazaitis, K., Mohamed, T., Nakashole, N., Platanios, E., Ritter, A., Samadi, M., Settles, B., Wang, R., Wijaya, D., Gupta, A., Chen, X., Saparov, A., Greaves, M., and Welling, J. (2015). Never-Ending Learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*, Austin, États-Unis.

Moro, A. and Navigli, R. (2015). SemEval-2015 Task 13 : Multilingual All-Words Sense Disambiguation and Entity Linking. In Nakov, P., Zesch, T., Cer, D., and Jurgens, D., editors, *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 288–297, Denver, États-Unis. Association for Computational Linguistics.

Moro, A., Raganato, A., and Navigli, R. (2014). Entity Linking meets Word Sense Disambiguation : a Unified Approach. *Transactions of the Association for Computational Linguistics*, 2 :231–244.

Moskvoretskii, V., Neminova, E., Lobanova, A., Panchenko, A., and Nikishina, I. (2024). TaxoLLaMA : WordNet-based Model for Solving Multiple Lexical Semantic Tasks. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 2331–2350, Bangkok, Thaïlande. Association for Computational Linguistics.

Mu, Y., Wu, B. P., Thorne, W., Robinson, A., Aletras, N., Scarton, C., Bontcheva, K., and Song, X. (2024). Navigating Prompt Complexity for Zero-Shot Classification : A Study of Large Language Models in Computational Social Science. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12074–12086, Turin, Italie. ELRA and ICCL.

Namer, F., Barque, L., Bonami, O., Haas, P., Hathout, N., and Tribout, D. (2019). Demonette2 - Une base de données dérivationnelle du français à grande échelle : premiers résultats. In Morin, E., Rosset, S., and Zweigenbaum, P., editors, *Conférence sur le Traitement Automatique des Langues Naturelles (TALN) - PFIA*, pages 233–244, Toulouse, France. ATALA.

Nasr, A. and Béchet, F. (2009). Analyse syntaxique en dépendances de l'oral spontané. In *Actes de la 16ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*, pages 21–30, Senlis, France. ATALA.

Nasr, A., Béchet, F., Rey, J.-F., Favre, B., and Le Roux, J. (2011). MACAON An NLP Tool Suite for Processing Word Lattices. In *Proceedings of the ACL-HLT 2011 System Demonstrations*, pages 86–91, Portland, Oregon. Association for Computational Linguistics.

Navigli, R. (2009). Word Sense Disambiguation : A survey. *ACM Comput. Surv.*, 41(2).

Navigli, R., Bevilacqua, M., Conia, S., Montagnini, D., and Cecconi, F. (2021). Ten Years of BabelNet : A Survey. In Zhou, Z.-H., editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4559–4567, Montreal, Quebec, Canada. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Navigli, R., Jurgens, D., and Vannella, D. (2013). SemEval-2013 Task 12 : Multilingual Word Sense Disambiguation. In Manandhar, S. and Yuret, D., editors, *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2 : Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231, Atlanta, États-Unis. Association for Computational Linguistics.

Navigli, R. and Ponzetto, S. P. (2012). BabelNet : The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193 :217–250.

Ng, H. T. and Lee, H. B. (1996). Integrating Multiple Knowledge Sources to Disambiguate Word Sense : An Exemplar-Based Approach. In *34th Annual Meeting of the Association for Computational Linguistics*, pages 40–47, Santa Cruz, États-Unis. Association for Computational Linguistics.

Nguyen, D. Q., Nguyen, T. D., Nguyen, D. Q., and Phung, D. (2018). A Novel Embedding Model for Knowledge Base Completion Based on Convolutional Neural Network. In Walker, M., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 2 (Short Papers)*, pages 327–333, La Nouvelle-Orléans, États-Unis. Association for Computational Linguistics.

Nickel, M., Tresp, V., and Kriegel, H.-P. (2011). A Three-Way Model for Collective Learning on Multi-Relational Data. In Getoor, L. and Scheffer, T., editors, *Proceedings of the 28th*

International Conference on Machine Learning (ICML-11), ICML '11, pages 809–816, New York, États-Unis. ACM.

Niles, I. and Pease, A. (2001). Towards a standard upper ontology. In *Proceedings of the International Conference on Formal Ontology in Information Systems - Volume 2001*, FOIS '01, page 2–9, New York, États-Unis. Association for Computing Machinery.

Oliver, A. and Climent, S. (2012). Parallel Corpora for WordNet Construction : Machine Translation vs. Automatic Sense Tagging. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, pages 110–121, Berlin, Heidelberg. Springer Berlin Heidelberg.

Ooi, V. B. Y. (1998). *Computer corpus lexicography / Vincent B.Y. Ooi.* Edinburgh textbooks in empirical linguistics. Edinburgh University Press, Edimbourg, Royaume-Uni.

Opitz, J., Wein, S., and Schneider, N. (2025). Natural Language Processing RELIES on Linguistics. *Computational Linguistics*, 51(3) :1009–1032.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The PageRank Citation Ranking : Bringing Order to the Web. Technical Report 1999-66, Stanford InfoLab. Previous number = SIDL-WP-1999-0120.

Panchenko, A., Naets, H., Brouwers, L., Romanov, P., and Fairon, C. (2013). Search and Visualization of Semantically Related Words (Recherche et visualisation de mots sémantiquement liés) [in French]. In Morin, E. and Estève, Y., editors, *Proceedings of TALN 2013 (Volume 2 : Short Papers)*, pages 747–754, Les Sables d'Olonne, France. ATALA.

Pasini, T., Raganato, A., and Navigli, R. (2021). XL-WSD : An Extra-Large and Cross-Lingual Evaluation Framework for Word Sense Disambiguation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15) :13648–13656.

Paulheim, H. (2017). Knowledge graph refinement : A survey of approaches and evaluation methods. *Semant. Web*, 8(3) :489–508.

Paulheim, H. and Bizer, C. (2014). Improving the Quality of Linked Data Using Statistical Distributions. *Int. J. Semant. Web Inf. Syst.*, 10(2) :63–86.

Paveau, M.-A. and Sarfati, G.-E. (2003). *Les grandes théories de la linguistique : de la grammaire comparée à la pragmatique / Marie-Anne Paveau, Georges-Elia Sarfati (dir.).* Collection U Lettres-Linguistique. Armand Colin, Paris.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe : Global Vectors for Word Representation. In Moschitti, A., Pang, B., and Daelemans, W., editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep Contextualized Word Representations. In Walker, M., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, États-Unis. Association for Computational Linguistics.

Peters, M. E., Neumann, M., Logan, R., Schwartz, R., Joshi, V., Singh, S., and Smith, N. A. (2019). Knowledge Enhanced Contextual Word Representations. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 43–54, Hong Kong, Chine. Association for Computational Linguistics.

Petrov, A., Venant, A., Lareau, F., Lepage, Y., and Langlais, P. (2025). ALF : Un jeu de données d'analogies françaises à grain fin pour l'évaluation de la connaissance lexicale des grands modèles de langue. In *Actes de CORIA-TALN-RJCRI-RECITAL 2025. Actes des 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : articles scientifiques originaux*, pages 22–49, Marseille, France. ATALA.

Pilehvar, M. T. and Camacho-Collados, J. (2019). WiC : the Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, États-Unis. Association for Computational Linguistics.

Ploux, S. and Victorri, B. (1998). Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes. *Revue TAL*, 39 :161–182.

Plu, J., Cousot, K., Lafourcade, M., Troncy, R., and Rizzo, G. (2018). JeuxDeLiens : Word Embeddings and Path-Based Similarity for Entity Linking using the French JeuxDeMots Lexical Semantic Network. In Sébillot, P. and Claveau, V., editors, *Actes de la Conférence TALN. Volume 1 - Articles longs, articles courts de TALN*, pages 529–538, Rennes, France. ATALA.

Polguère, A. (2018). A Lexicographic Approach to the Study of Copolysemy Relations *. *Russian Journal of Linguistics = Vestnik Rossijskogo universiteta družby narodov. Seriâ Lingvistika*, 22(4) :788 – 820.

Polguère, A. (2016). *Lexicologie et sémantique lexicale : Notions fondamentales / Alain Polguère*. Paramètres. Les Presses de l'Université de Montréal, Montréal, 3e édition edition.

Postma, M., Izquierdo, R., Agirre, E., Rigau, G., and Vossen, P. (2016). Addressing the MFS Bias in WSD systems. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1695–1700, Portorož, Slovénie. European Language Resources Association (ELRA).

Pradhan, S., Loper, E., Dligach, D., and Palmer, M. (2007). SemEval-2007 Task-17 : English Lexical Sample, SRL and All Words. In Agirre, E., Màrquez, L., and Wicentowski, R., editors, *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92, Prague, République Tchèque. Association for Computational Linguistics.

Prost, J.-P. (2022). Integrating a Phrase Structure Corpus Grammar and a Lexical-Semantic Network : the HOLINET Knowledge Graph. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 613–622, Marseille, France. European Language Resources Association.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza : A Python Natural Language Processing Toolkit for Many Human Languages. In Celikyilmaz, A. and Wen, T.-H., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, pages 101–108, En distanciel. Association for Computational Linguistics.

Qwen, :, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., and Qiu, Z. (2025). Qwen2.5 Technical Report.

Raganato, A., Camacho-Collados, J., and Navigli, R. (2017). Word Sense Disambiguation : A Unified Evaluation Framework and Empirical Comparison. In Lapata, M., Blunsom, P., and Koller, A., editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 1, Long Papers*, pages 99–110, Valence, Espagne. Association for Computational Linguistics.

Ramachandran, P., Zoph, B., and Le, Q. V. (2018). Searching for Activation Functions. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, Vancouver, Canada. Workshop Track.

Ramadier, L. and Lafourcade, M. (2018). Utilisation d'une base de connaissances de spécialité et de sens commun pour la simplification de comptes-rendus radiologiques (Radiological text simplification using a general knowledge base). In Sébillot, P. and Claveau, V., editors, *Actes de la Conférence TALN. Volume 1 - Articles longs, articles courts de TALN*, pages 447–454, Rennes, France. ATALA.

Randriatsitohaina, T. and Hamon, T. (2020). Identification des problèmes d'annotation pour l'extraction de relations (Identification of annotation problem for the relation extraction). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles*, pages 323–331, Nancy, France. ATALA et AFCP.

Rauzy, S. and Blache, P. (2007). Un lexique syntaxique des verbes du français : VfrLPL. 7 pages.

Reiter, E. (2025). We Should Evaluate Real-World Impact. *Computational Linguistics*, 51(4) :1419–1431.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'95, page 448–453, San Francisco, États-Unis. Morgan Kaufmann Publishers Inc.

Riezler, S., King, T. H., Kaplan, R. M., Crouch, R., Maxwell III, J. T., and Johnson, M. (2002). Parsing the Wall Street Journal using a Lexical-Functional Grammar and Discriminative Estimation Techniques. In Isabelle, P., Charniak, E., and Lin, D., editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 271–278, Philadelphie, États-Unis. Association for Computational Linguistics.

Riloff, E. and Shepherd, J. (1997). A Corpus-Based Approach for Building Semantic Lexicons. In *Second Conference on Empirical Methods in Natural Language Processing*.

Rogers, A. and Luccioni, A. S. (2024). Position : key claims in LLM research have a long tail of footnotes. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Rohatgi, S. (2022). ACL Anthology Corpus with Full Text. Github.

Sabatier, P. and Le Pesant, D. (2013). Les dictionnaires électroniques de Jean Dubois et Françoise Dubois-Charlier et leur exploitation en TAL. In *Ressources lexicales : contenu, évaluation, utilisation, évaluation.*, volume 30 of *Lingvisticæ Investigationes Supplementa*, pages 153–186. John Benjamins Publishing.

Sagot, B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *7th international conference on Language Resources and Evaluation (LREC 2010)*, La Valette, Malte.

Sagot, B. (2013). Construction de ressources lexicales pour le traitement automatique des langues. In *Ressources lexicales : contenu, évaluation, utilisation, évaluation.*, volume 30 of *Lingvisticæ Investigationes Supplementa*, pages 217–254. John Benjamins Publishing.

Sagot, B. (2019). Développement d'un lexique morphologique et syntaxique de l'ancien français (Development of a morphological and syntactic lexicon of Old French). In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Volume II : Articles courts*, pages 265–274, Toulouse, France. ATALA.

Sagot, B. and Fišer, D. (2008). Building a free French wordnet from multilingual resources. In *OntoLex*, Marrakech, Maroc.

Sagot, B., Fort, K., Adda, G., Mariani, J., and Lang, B. (2011). Un turc mécanique pour les ressources linguistiques : critique de la myriadisation du travail parcellisé. In *TALN'2011 - Traitement Automatique des Langues Naturelles*, Montpellier, France.

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., and Chadha, A. (2025). A Systematic Survey of Prompt Engineering in Large Language Models : Techniques and Applications.

Sajous, F., Hathout, N., and Calderone, B. (2013). GLÁFF, un Gros Lexique Á tout Faire du Français. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, pages 285–298, Les Sables d'Olonne, France. ATALA.

Sakajo, H., Ide, Y., Vasselli, J., Sakai, Y., Tian, Y., Kamigaito, H., and Watanabe, T. (2025). Dictionaries to the Rescue : Cross-Lingual Vocabulary Transfer for Low-Resource Languages Using Bilingual Dictionaries. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T., editors, *Findings of the Association for Computational Linguistics : ACL 2025*, pages 25963–25976, Vienne, Austriche. Association for Computational Linguistics.

Saussure, F. d., Bally, C., Sechehaye, A., Riedlinger, A., De Mauro, T., and Louis-Jean, C. (1995). *Cours de linguistique générale / Ferdinand de Saussure ; publié par Charles Bailly [i.e. Bally] et Albert Séchehaye avec la collaboration de Albert Riedlinger édition critique préparée par Tullio De Mauro postface de Louis-Jean Calvet.* Grande bibliothèque Payot. Payot, Paris, [reproduction en fac-similé] edition.

Scarlini, B., Pasini, T., and Navigli, R. (2020). SensEmBERT : Context-Enhanced Sense Embeddings for Multilingual Word Sense Disambiguation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05) :8758–8765.

Schalchli, G. (2022). Lexique4linguists. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

Schlichtkrull, M., Kipf, T. N., Bloem, P., van den Berg, R., Titov, I., and Welling, M. (2018). Modeling Relational Data with Graph Convolutional Networks. In Gangemi, A., Navigli, R., Vidal, M.-E., Hitzler, P., Troncy, R., Hollink, L., Tordai, A., and Alam, M., editors, *The Semantic Web*, pages 593–607. Springer International Publishing.

Schneider, E. W. (1973). Course Modularization Applied : The Interface System and Its Implications For Sequence Control and Data Analysis. In *Association for the Development of Instructional Systems (ADIS)*, Chicago, États-Unis.

Schneider, P., Schopf, T., Vladika, J., Galkin, M., Simperl, E., and Matthes, F. (2022). A Decade of Knowledge Graphs in Natural Language Processing : A Survey. In He, Y., Ji, H., Li, S., Liu, Y., and Chang, C.-H., editors, *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, pages 601–614, En distanciel. Association for Computational Linguistics.

Scozzafava, F., Maru, M., Brignone, F., Torrisi, G., and Navigli, R. (2020). Personalized PageRank with Syntagmatic Information for Multilingual Word Sense Disambiguation. In Celikyilmaz, A. and Wen, T.-H., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, pages 37–46, En distanciel. Association for Computational Linguistics.

Segonne, V. (2021). *French verb sense disambiguation.* Thèse de doctorat, Université Paris Cité.

Segonne, V., Candito, M., and Crabbé, B. (2019). Using Wiktionary as a resource for WSD : the case of French verbs. In Dobnik, S., Chatzikyriakidis, S., and Demberg, V., editors, *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 259–270, Göteborg, Suède. Association for Computational Linguistics.

Sérasset, G. (2015). DBnary : Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF. *Semantic Web – Interoperability, Usability, Applicability*, 6(4) :355–361.

Shen, L., Tan, W., Chen, S., Chen, Y., Zhang, J., Xu, H., Zheng, B., Koehn, P., and Khashabi, D. (2024). The Language Barrier : Dissecting Safety Challenges of LLMs in Multilingual Contexts. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Findings of the Association for Computational Linguistics : ACL 2024*, pages 2668–2680, Bangkok, Thaïlande. Association for Computational Linguistics.

Shu, D., Chen, T., Jin, M., Zhang, C., Du, M., and Zhang, Y. (2024). Knowledge Graph Large Language Model (KG-LLM) for Link Prediction. In Nguyen, V. and Lin, H., editors, *Asian Conference on Machine Learning (ACML)*, volume 260 of *Proceedings of Machine Learning Research*, pages 143–158, Hanoi, Vietnam. PMLR.

Singh, C., Inala, J. P., Galley, M., Caruana, R., and Gao, J. (2024). Rethinking Interpretability in the Era of Large Language Models.

Sinha, A., Ollinger, S., and Constant, M. (2022). Word Sense Disambiguation of French Lexicographical Examples Using Lexical Networks. In Ustalov, D., Gao, Y., Panchenko, A., Valentino, M., Thayaparan, M., Nguyen, T. H., Penn, G., Ramesh, A., and Jana, A., editors, *Proceedings of TextGraphs-16 : Graph-based Methods for Natural Language Processing*, pages 70–76, Gyeongju, République de Corée. Association for Computational Linguistics.

Snyder, B. and Palmer, M. (2004). The English all-words task. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43, Barcelone, Espagne. Association for Computational Linguistics.

Spearman, C. (2010). The proof and measurement of association between two things. *International Journal of Epidemiology*, 39(5) :1137–1150.

Sri Nurdiati, S. and Hoede, C. (2008). *25 years development of knowledge graph theory : the results and the challenge*. Number 2/1876 in Memorandum. Université de Twente, Pays-Bas.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., and Varga, D. (2006). The JRC-Acquis : A Multilingual Aligned Parallel Corpus with 20+ Languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Gène, Italie.

Strnadová, J. and Sagot, B. (2011). Construction d'un lexique des adjectifs dénominaux (Construction of a lexicon of denominal adjectives). In *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, pages 67–72, Montpellier, France. ATALA.

Suchanek, F. M., Abiteboul, S., and Senellart, P. (2011). PARIS : probabilistic alignment of relations, instances, and schema. *Proc. VLDB Endow.*, 5(3) :157–168.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). Yago : A Core of Semantic Knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, page 697–706, New York, États-Unis. Association for Computing Machinery.

Sumanathilaka, D. K., Micallef, N., and Hough, J. (2024). Can LLMs assist with Ambiguity ? A Quantitative Evaluation of various Large Language Models on Word Sense Disambiguation. In Mitkov, R., Ezzini, S., Ranasinghe, T., Ezeani, I., Khallaf, N., Acarturk, C., Bradbury, M., El-Haj, M., and Rayson, P., editors, *Proceedings of the First International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security*, pages 97–108, Lancaster, Royaume-Uni. International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security.

Sun, Z., Deng, Z., Nie, J., and Tang, J. (2019). RotatE : Knowledge Graph Embedding by Relational Rotation in Complex Space. In *7th International Conference on Learning Representations, ICLR 2019*, La Nouvelle-Orléans, États-Unis.

Tack, A., François, T., Ligozat, A.-L., and Fairon, C. (2016). Modèles adaptatifs pour prédire automatiquement la compétence lexicale d'un apprenant de français langue étrangère (Adaptive models for automatically predicting the lexical competence of French as a foreign language learners). In Danlos, L. and Hamon, T., editors, *Actes de la conférence conjointe JEP-TALN-RECITAL 2016. volume 2 : TALN (Articles longs)*, pages 221–234, Paris, France. AFCP - ATALA.

Taghipour, K. and Ng, H. T. (2015). One Million Sense-Tagged Instances for Word Sense Disambiguation and Induction. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 338–344, Pékin, Chine. Association for Computational Linguistics.

Taghizadeh, N. and Faili, H. (2016). Automatic Wordnet Development for Low-Resource Languages Using Cross-Lingual WSD. *J. Artif. Int. Res.*, 56(1) :61–87.

Tanguy, L., Brunet, P., and Ferret, O. (2019). Comparaison qualitative et extrinsèque d'analyseurs syntaxiques du français : confrontation de modèles distributionnels sur un corpus spécialisé (Extrinsic evaluation of French dependency parsers on a specialised corpus : comparison of distributional thesauri). In Morin, E., Rosset, S., and Zweigenbaum, P., editors, *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Volume I : Articles longs*, pages 39–54, Toulouse, France. ATALA.

Tanguy, L., Fabre, C., Hathout, N., and Ho-Dac, L.-M. (2025). Embeddings, topic models, LLM : un air de famille. In *Actes de CORIA-TALN-RJCRI-RECITAL 2025. Actes des 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : articles scientifiques originaux*, pages 295–312, Marseille, France. ATALA.

Tanguy, L. and Hathout, N. (2002). Webaffix : un outil d'acquisition morphologique dérivationnelle à partir du Web. In Pierrel, J.-M., editor, *Actes de la 9eConférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN-2002)*, pages 245–254, Nancy, France. ATALA.

Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., Rouillard, L., Mesnard, T., Cideron, G., bastien Grill, J., Ramos, S., Yvinec, E., Casbon, M., Pot, E., Penchev, I., Liu, G., Visin, F., Kenealy, K., Beyer, L., Zhai, X., Tsitsulin, A., Busa-Fekete, R., Feng, A., Sachdeva, N., Coleman, B., Gao, Y., Mustafa, B., Barr, I., Parisotto, E., Tian, D., Eyal, M., Cherry, C., Peter, J.-T., Sinopalnikov, D., Bhupatiraju, S., Agarwal, R., Kazemi, M., Malkin, D., Kumar, R., Vilar, D., Brusilovsky, I., Luo, J., Steiner, A., Friesen, A., Sharma, A., Sharma, A., Gilady, A. M., Goedeckemeyer, A., Saade, A., Feng, A., Kolesnikov, A., Bendebury, A., Abdagic, A., Vadi, A., György, A., Pinto, A. S., Das, A., Bapna, A., Miech, A., Yang, A., Paterson, A., Shenoy, A., Chakrabarti, A., Piot, B., Wu, B., Shahriari, B., Petrini, B., Chen, C., Lan, C. L., Choquette-Choo, C. A., Carey, C., Brick, C., Deutsch, D., Eisenbud, D., Cattle, D., Cheng, D., Paparas, D., Sreepathihalli, D. S., Reid, D., Tran, D., Zelle, D., Noland, E., Huizenga, E., Kharitonov, E., Liu, F., Amirkhanyan, G., Cameron, G., Hashemi, H., Klimczak-Plucińska, H., Singh, H., Mehta, H., Lehri, H. T., Hazimeh, H., Ballantyne, I., Szpektor, I., Nardini, I., Pouget-Abadie, J., Chan, J., Stanton, J., Wieting, J., Lai, J., Orbay, J., Fernandez, J., Newlan, J., yeong Ji, J., Singh, J., Black, K., Yu, K., Hui, K., Vodrahalli, K., Greff, K., Qiu, L., Valentine, M., Coelho, M., Ritter, M., Hoffman, M., Watson, M., Chaturvedi, M., Moynihan, M., Ma, M., Babar, N., Noy, N., Byrd, N., Roy, N., Momchev, N., Chauhan, N., Sachdeva, N., Bunyan, O., Botarda, P., Caron, P., Rubenstein, P. K., Culliton, P., Schmid, P., Sessa, P. G., Xu, P., Stanczyk, P., Tafti, P., Shivanna, R., Wu, R., Pan, R., Rokni, R., Willoughby, R., Vallu, R., Mullins, R., Jerome, S., Smoot, S., Girgin, S., Iqbal, S., Reddy, S., Sheth, S., Põder, S., Bhatnagar, S., Panyam, S. R., Eiger, S., Zhang, S., Liu, T., Yacovone, T., Liechty, T., Kalra, U., Evci, U., Misra, V., Roseberry, V., Feinberg, V., Kolesnikov, V., Han, W., Kwon, W., Chen, X., Chow, Y., Zhu, Y., Wei, Z., Egyed, Z., Cotruta, V., Giang, M., Kirk, P., Rao, A., Black, K., Babar, N., Lo, J., Moreira, E., Martins, L. G., Sanseviero, O., Gonzalez, L., Gleicher, Z., Warkentin, T., Mirrokni, V., Senter, E., Collins,

E., Barral, J., Ghahramani, Z., Hadsell, R., Matias, Y., Sculley, D., Petrov, S., Fiedel, N., Shazeer, N., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Alayrac, J.-B., Anil, R., Dmitry, Lepikhin, Borgeaud, S., Bachem, O., Joulin, A., Andreev, A., Hardin, C., Dadashi, R., and Hussenot, L. (2025). Gemma 3 Technical Report.

Tellier, I., Eshkol-Taravella, I., Dupont, Y., and Wang, I. (2014). Can we chunk well with bad POS labels ? (Peut-on bien chunker avec de mauvaises étiquettes POS ?) [in French]. In Blache, P., Béchet, F., and Bigi, B., editors, *Proceedings of TALN 2014 (Volume 1 : Long Papers)*, pages 125–136, Marseille, France. ATALA.

Teru, K. K., Denis, E. G., and Hamilton, W. L. (2020). Inductive relation prediction by subgraph reasoning. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20, Vienne, Autriche. JMLR.org.

Todirascu, A., Cargill, M., and Francois, T. (2019). PolylexFLE : une base de données d'expressions polylexicales pour le FLE (PolylexFLE : a database of multiword expressions for French L2 language learning). In Morin, E., Rosset, S., and Zweigenbaum, P., editors, *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Volume I : Articles longs*, pages 143–156, Toulouse, France. ATALA.

Tolone, E. (2009). Les tables du Lexique-Grammaire au format TAL. In *Actes de MajecSTIC 2009*, page electronic version (8 pp.), Avignon, France.

Toutanova, K. and Chen, D. (2015). Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, pages 57–66, Pékin, Chine. Association for Computational Linguistics.

Tran, M. and Maurel, D. (2006). Prolexbase - Un dictionnaire relationnel multilingue de noms propres. *Trait. Autom. des Langues*, 47(3) :115–139.

Trouilleux, F. (2012). Le DM, a French Dictionary for NooJ. In Kristina Vučković, B. B. and Silberztein, M., editors, *Automatic Processing of Various Levels of Linguistic Phenomena : Selected Papers from the NooJ 2011 International Conference*, pages 16–28. Cambridge Scholars Publishing.

Trouillon, T., Welbl, J., Riedel, S., Gaussier, E., and Bouchard, G. (2016). Complex Embeddings for Simple Link Prediction. In Balcan, M. F. and Weinberger, K. Q., editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2071–2080, New York, États-Unis. PMLR.

Urieli, A. and Tanguy, L. (2013). L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur Talismane). In Morin, E. and Estève, Y., editors, *Proceedings of TALN 2013 (Volume 1 : Long Papers)*, pages 188–201, Les Sables d'Olonne, France. ATALA.

Van den Eynde, K. and Mertens, P. (2006). Le dictionnaire de valence DICOVALENCE : manuel d'utilisation.

Van den Eynde, K. and Mertens, P. (2010). Le dictionnaire de valence DICOVALENCE : manuel d'utilisation version 2.0.

Vashishth, S., Sanyal, S., Nitin, V., and Talukdar, P. P. (2020). Composition-based Multi-Relational Graph Convolutional Networks. In *8th International Conference on Learning Representations, ICLR 2020*, Addis-Abeba, Éthiopie.

Vasilescu, F., Langlais, P., and Lapalme, G. (2004). Evaluating Variants of the Lesk Approach for Disambiguating Words. In Lino, M. T., Xavier, M. F., Ferreira, F., Costa, R., and Silva, R., editors, *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbonne, Portugal. European Language Resources Association (ELRA).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, États-Unis. Curran Associates Inc.

Vial, L., Lecouteux, B., and Schwab, D. (2019). Sense Vocabulary Compression through the Semantic Knowledge of WordNet for Neural Word Sense Disambiguation. In Vossen, P. and Fellbaum, C., editors, *Proceedings of the 10th Global Wordnet Conference*, pages 108–117, Wroclaw, Pologne. Global Wordnet Association.

Vossen, P., editor (1998). *EuroWordNet : a multilingual database with lexical semantic networks*. Kluwer Academic Publishers, États-Unis.

Wahle, J. P., Ruas, T., Xu, Y., and Gipp, B. (2024). Paraphrase Types Elicit Prompt Engineering Capabilities. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11004–11033, Miami, États-Unis. Association for Computational Linguistics.

Walsh, B., Mohamed, S. K., and Nováček, V. (2020). BioKG : A Knowledge Graph for Relational Learning On Biological Data. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 3173–3180, New York, NY, USA. Association for Computing Machinery.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). GLUE : A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In Linzen, T., Chrupała, G., and Alishahi, A., editors, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP : Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Wang, M. and Wang, Y. (2020). A Synset Relation-enhanced Framework with a Try-again Mechanism for Word Sense Disambiguation. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6229–6240, En distanciel. Association for Computational Linguistics.

Wang, Q., Mao, Z., Wang, B., and Guo, L. (2017). Knowledge Graph Embedding : A Survey of Approaches and Applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12) :2724–2743.

Wang, Z., Zhang, J., Feng, J., and Chen, Z. (2014). Knowledge Graph Embedding by Translating on Hyperplanes. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI'14, page 1112–1119, Québec City, Québec, Canada. AAAI Press.

Weaver, W. (1955). Translation. In Locke, W. N. and Booth, A. D., editors, *Machine Translation of Languages : Fourteen Essays*, pages 15–23. Technology Press of the Massachusetts Institute of Technology and John Wiley & Sons, Inc., Cambridge, MA and New York. Written July 15, 1949.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, États-Unis. Curran Associates Inc.

Xiao, Y., Zhang, S., Zhou, H., Li, M., Yang, H., and Zhang, R. (2024). FuseLinker : Leveraging LLM's pre-trained text embeddings and domain knowledge to enhance GNN-based link prediction on biomedical knowledge graphs. *Journal of Biomedical Informatics*, 158 :104730.

Xiong, W., Hoang, T., and Wang, W. Y. (2017). DeepPath : A Reinforcement Learning Method for Knowledge Graph Reasoning. In Palmer, M., Hwa, R., and Riedel, S., editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 564–573, Copenhague, Danemark. Association for Computational Linguistics.

Xu, J., Kim, S., Song, M., Jeong, M., Kim, D., Kang, J., Rousseau, J. F., Li, X., Xu, W., Torvik, V. I., Bu, Y., Chen, C., Ebeid, I. A., Li, D., and Ding, Y. (2020). Building a PubMed knowledge graph. *Scientific Data*, 7(1) :205.

Yang, B., Yih, W., He, X., Gao, J., and Deng, L. (2015). Embedding Entities and Relations for Learning and Inference in Knowledge Bases. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings*, San Diego, États-Unis.

Yarowsky, D., Cucerzan, S., Florian, R., Schafer, C., and Wicentowski, R. (2001). The John Hopkins SENSEVAL-2 System Descriptions. In Preiss, J. and Yarowsky, D., editors, *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 163–166, Toulouse, France. Association for Computational Linguistics.

Yu, M. and Dredze, M. (2014). Improving Lexical Embeddings with Semantic Knowledge. In Toutanova, K. and Wu, H., editors, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, pages 545–550, Baltimore, États-Unis. Association for Computational Linguistics.

Zeng, K., Li, C., Hou, L., Li, J., and Feng, L. (2021). A comprehensive survey of entity alignment for knowledge graphs. *AI Open*, 2 :1–13.

Zhang, K., Choi, Y., Song, Z., He, T., Wang, W. Y., and Li, L. (2024). Hire a Linguist ! : Learning Endangered Languages in LLMs with In-Context Linguistic Descriptions. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Findings of the Association for Computational Linguistics : ACL 2024*, pages 15654–15669, Bangkok, Thaïlande. Association for Computational Linguistics.

Zhao, B., Sun, J., Xu, B., Lu, X., Li, Y., Yu, J., Liu, M., Zhang, T., Chen, Q., Li, H., Hou, L., and Li, J. (2022). EDUKG : a Heterogeneous Sustainable K-12 Educational Knowledge Graph.