



Annotation semi-automatique en sens des noms communs du
corpus français FrSemCor avec le modèle AMuSe-WSD

Mémoire de Master 1 Langue et Informatique

Présenté par :

Anîsa AMOURAT et Assia BOUZNAD

Sous la direction de :

Karën FORT (Sorbonne Université, LORIA)
Hee-Soo CHOI (Université de Lorraine, ATILF/LORIA)

Table des matières

Introduction	1
1 Définitions et État de l'art	3
1.1 Description de la tâche de WSD	3
1.2 État de l'art	4
1.2.1 FlauBERT : un modèle de langue évalué sur la tâche de désambiguïsation lexicale	4
1.2.2 Wiktionnaire : un inventaire de sens pour les annotations en sens du français	5
1.2.3 SemEval2013 : un corpus textuel d'annotation en sens de qualité pour la désambiguïsation lexicale en français	6
2 Corpus et Outils de recherche	9
2.1 Corpus	9
2.1.1 Description du corpus SemEval2013	9
2.1.2 Description du corpus FrSemCor : un corpus français annoté avec des <i>supersenses</i>	12
2.1.3 L'annotation en <i>Supersenses</i>	15
2.2 Ressources lexicographiques	16
2.2.1 WordNet : un inventaire de sens pour l'anglais	16
2.2.2 BabelNet : un inventaire de sens multilingue	18
2.3 Description du modèle AMuSe-WSD	19
3 Évaluation du modèle AMuSe-WSD sur le corpus SemEval2013	23
3.1 Récupération des données du corpus SemEval2013 obtenu par le modèle AMuSe-WSD	23
3.2 Annotation automatique du corpus SemEval2013 avec AMuSe-WSD	25
3.2.1 Comparaison de la tokenisation et de l'étiquetage en partie du discours	25
3.2.2 Comparaison des annotations en sens	26
4 Utilisation du modèle AMuSe-WSD sur le corpus FrSemCor	29
4.1 Statistiques sur les annotations en <i>supersenses</i> des corpus SemEval2013 et FrSemCor	29
4.2 Annotations manuelles en sens sur le corpus FrSemCor	31
4.3 Observation des annotations manuelles effectuées sur le corpus FrSemCor	33
4.3.1 Calcul de l'accord inter-annotateurs	33
4.3.2 Comparaison des annotations générées par AMuSe-WSD avec les annotations <i>gold</i> sur le corpus FrSemCor	35
Conclusion et Perspectives	39
Intégration d'une détection des <i>supersenses</i> au code source du modèle AMuSe-WSD	40
Amélioration de la tokenisation par AMuSe-WSD	41

TABLE DES MATIÈRES

Augmentation des données d'entraînement du modèle	41
Annexes	43
Tableau d'annotations manuelles sur le corpus FrSemCor	43

Table des figures

1.1	Tableau des performances finales sur les tâches de FLUE (Kitaev et al., 2019).(Constant et al., 2013).(Eisenschlos et al., 2019).(Chen et al., 2017).(Martin et al., 2020).(Segonne et al., 2019)(Tableau 2 de (Le et al., 2020))	5
1.2	Taux d'ambiguïté des verbes, en anglais usuel ensembles, ainsi que dans l'ensemble de formation (<i>Wiktionnaire</i>) et l'ensemble d'évaluation (FSE) français. AMBIG_trainSI correspond à utiliser pour le nombre de sens l'inventaire des sens dans le corpus d'entraînement correspondant, tandis qu' AMBIG_fullSI correspond à utiliser l'inventaire des sens complet. (Tableau 3 de l'article de (Segonne et al., 2019))	6
2.1	Exemple d'annotation manuelle dans le corpus SemEval2013-fr	10
2.2	Répartition des étiquettes de partie du discours dans SemEval2013	10
2.3	Extraits des noms ne possédant pas d'annotation en sens	11
2.4	Diagramme représentant les proportions des noms possédant une annotation en sens et ceux qui n'en possèdent pas	12
2.5	Exemple d'affichage des <i>synsets</i> dans WordNet	17
2.6	Présentation de l'interface de BabelNet	18
2.7	Présentation de l'interface du modèle AMuSe-WSD	19
2.8	Représentation des <i>Words Embeddings</i> (Image from polakowo.io)	20
3.1	Schéma récapitulatif de notre utilisation du modèle AMuSe-WSD sur le corpus SemEval2013	23
3.2	Récupération des proportions de chaque catégorie grammaticale du corpus SemEval2013 à partir de l'API du modèle AMuSe-WSD	24
3.3	Représentation des proportions des cas dans lesquels les expressions polylexicales (c.à.d. les noms écrits avec des <i>underscores</i>) apparaissent uniquement dans les données d' AMuSe-WSD , uniquement dans les données du SemEval2013 ou alors dans les deux jeux de données.	25
3.4	Représentation des proportions des cas dans lesquels les noms apparaissent uniquement dans les données d' AMuSe-WSD , uniquement dans les données du SemEval2013 ou alors dans les deux jeux de données.	26
3.5	Représentation graphique des <i>matchs</i> de la comparaison des <i>synsets</i>	27
4.1	Proportions de <i>supersenses</i> trouvés dans le SemEval2013	30
4.2	Extrait du corpus FrSemCor au format CoNLL-U (Barque et al., 2020) . .	32
4.3	Extrait de nos annotations manuelles du corpus FrSemCor	33
4.4	Formule permettant de calculer le score Kappa de Cohen, où $\text{Pr}(a)$ est la proportion d'accord observé entre annotateurs (ou l'accord observé) et $\text{Pr}(e)$ la probabilité d'accord aléatoire.	34

TABLE DES FIGURES

4.5	Tableau d'interprétation du Kappa de Cohen selon Wong et al. (2021). Les ordres de grandeur proposés ne font pas consensus dans la communauté scientifique, en raison de l'influence du nombre de catégories sur l'estimation : moins il y a de catégories, plus le κ est élevé.	34
4.6	Extrait de l'affichage des annotations AMuSe-WSD avec un exemple de l'identifiant BabelNet encadré en orange et de l'identifiant WordNet encadré en rose)	36
4.7	Extrait de notre tableau d'annotation avec les différents annotateurs et les correspondances ou non des <i>synsets</i>	36
4.8	Résultats des comparaisons faites entre les <i>synsets</i> BabelNet de nos annotations et ceux générés par AMuSe-WSD	37
4.9	Représentation graphique des <i>matchs</i> de la comparaison des <i>synsets</i> BabelNet	38

Liste des tableaux

2.1	<i>Supersenses</i> utilisés en annotation dans FrSemCor (à gauche), regroupés en classes plus générales (à droite). Les <i>supersenses</i> barrés sont les UBs qui n'ont pas été retenus, et les <i>supersenses</i> en gras sont ceux qui ont été ajoutés (cf. Tableau 1 de (Barque et al., 2020))	14
2.2	Exemples de relations lexicales dans WordNet	16
2.3	Résultats de la désambiguisation lexicale en anglais avec les scores F1 sur Senseval-2 (SE2), Senseval-3 (SE3), SemEval-2007 (SE07), SemEval-2013 (SE13), SemEval-2015 (SE15), ainsi que la concaténation de tous les ensembles de données (ALL). Nous incluons également les résultats sur les ensembles de données multilingues SemEval-2013 (SE13) et SemEval-2015 (SE15). (cf. Tableau 1 de l'article d'Orlando et al. (2021))	22
4.1	Distribution des 20 <i>supersenses</i> les plus fréquents dans le corpus annoté FrSemCor (= ceux utilisés pour plus de 50 tokens).(cf. Tableau 3 de Barque et al. (2020))	31

Remerciements

Nous souhaitons exprimer notre sincère reconnaissance à toutes les personnes qui ont contribué, de près ou de loin, à l'élaboration de ce mémoire.

Tout d'abord, nous tenons à remercier notre encadrante, Madame Hee-Soo CHOI, pour son encadrement bienveillant, ses précieux conseils et son soutien tout au long de ce travail. Ses remarques et suggestions nous ont permis d'approfondir nos réflexions et d'enrichir cette étude.

Nous remercions également notre directrice de mémoire, Madame Karen FÖRT, qui nous a permis de travailler sur ce projet et d'en apprendre encore un peu plus sur le domaine du Traitement Automatique des Langues.

Un grand merci à nos amis, Orphila, Alexis, Valérie et Clara et nos familles, dont les discussions et le soutien moral nous ont été précieux. Vos encouragements et vos remarques constructives nous ont aidé à surmonter nos moments de doute et ont permis à ce travail de voir le jour.

Anîsa AMOURAT et Assia BOUZNAD

Introduction

Dans le domaine étendu du Traitement Automatique des Langues (TAL), la désambiguïsation lexicale ou *Word Sense Disambiguation* (WSD), constitue un défi fondamental. Cette tâche consiste à déterminer le sens exact d'un mot en fonction du contexte dans lequel il apparaît. Les mots polysémiques, qui ont plusieurs sens, posent un problème particulier car leur interprétation dépend fortement du contexte d'utilisation. Par exemple, le terme « banque » peut désigner soit une institution financière, soit un lieu de repos, selon le contexte dans lequel il est employé.

La nécessité de désambiguïser les mots est omniprésente dans diverses applications du TAL, telles que la traduction automatique, l'analyse des sentiments ou encore la recherche d'information ([Navigli, 2009](#)). Une désambiguïsation précise est importante pour garantir la qualité et la précision des résultats générés par ces systèmes automatiques.

La tâche de désambiguïsation lexicale est généralement abordée comme un problème de classification où chaque occurrence d'un mot est attribuée à l'un de ses sens possibles. Pour accomplir cette tâche, deux éléments essentiels sont requis : un inventaire de sens, qui liste tous les sens possibles des mots, et des données annotées de qualité nécessaires pour l'entraînement et l'évaluation des modèles.

En plus des annotations en sens, il existe des annotations en *supersenses*. Tandis que les annotations en sens visent à préciser le sens exact des mots dans leur contexte, les annotations en *supersenses* regroupent les mots ou expressions en catégories plus générales, offrant une annotation de catégorie sémantique plutôt que du sens précis. Par exemple, le mot « souris » pourrait être associé au *supersense Artefact* pour désigner l'objet informatique, ou au *supersense Animal*, selon le contexte.

Cependant, un obstacle se présente par rapport à la rareté des données annotées en sens, en particulier pour les langues autres que l'anglais. En français, par exemple, il existe peu de ressources annotées de qualité pour entraîner et évaluer les systèmes de désambiguïsation lexicale. On retrouve notamment des données annotées en français, grâce à la traduction automatique, ce qui laisse matière à discuter sur la qualité de ces annotations.

Ce mémoire vise à explorer différentes approches pour améliorer la disponibilité des données annotées en sens pour le français. Nous nous proposons d'enrichir les données en créant un corpus annoté en sens manuellement en procédant comme suit :

- Évaluation d'un modèle, **AMuSe-WSD**, qui annote un texte en sens, avec un corpus annoté manuellement, **SemEval2013**.
- Annotation manuelle du corpus **FrSemCor**.
- Comparaison des annotations manuelles et automatiques générées par **AMuSe-WSD** du corpus **FrSemCor**.

INTRODUCTION

Répondre à ces objectifs permettra de combler une lacune importante dans les ressources disponibles pour le français en matière de WSD, ouvrant ainsi la voie à des avancées significatives dans le domaine du TAL pour les applications francophones.

Les efforts visent à fournir des données annotées de qualité, précieuses pour établir des modèles de WSD de référence pour les recherches futures.

Ce mémoire se structure en plusieurs parties : nous commencerons par introduire le sujet et définir le cadre de notre recherche. Ensuite, nous détaillerons la méthodologie employée pour enrichir une ressource, ici le **FrSemCor**, essentiel à la tâche de désambiguïsation lexicale. Enfin, nous partagerons nos expériences et réflexions tout au long de notre recherche, avant d'analyser les résultats obtenus.

Chapitre 1

Définitions et État de l'art

1.1 Description de la tâche de WSD

La tâche de levée d'ambiguïté lexicale (WSD, pour *Word Sense Disambiguation*) représente un défi dans le domaine du traitement automatique des langues (TAL) ([Bevilacqua et al., 2021](#)). Son objectif principal est de résoudre les ambiguïtés qui surgissent lorsque les mots ont plusieurs sens, en identifiant celui qui convient le mieux dans un contexte donné. Cette tâche est nécessaire pour de nombreuses applications du TAL comme par exemple, la traduction automatique, car la compréhension précise du sens des mots est essentielle pour interpréter correctement les textes et communiquer efficacement avec les utilisateurs.

Pour illustrer cette nécessité, prenons comme exemple le mot « moteur » : « Il a réparé le moteur de sa voiture. », ici, le mot « moteur » désigne un dispositif mécanique utilisé pour propulser le véhicule. En revanche, dans une phrase comme « Le moteur de recherche est financé par des subventions », le terme « moteur » est utilisé au sens figuré pour désigner le principal moteur ou facteur qui stimule l'activité de recherche. Cette variation illustre la difficulté des systèmes de désambiguïsation lexicale à distinguer le sens correct d'un mot.

Pour effectuer la désambiguïsation en sens lexicaux, les systèmes utilisent une variété de techniques et de ressources. En effet, cette tâche nécessite, d'une part, des techniques telles que des algorithmes de TAL pour analyser et interpréter les données textuelles ou encore des méthodes basées sur l'apprentissage automatique qui entraînent des modèles afin de déterminer les sens des mots dans un contexte donné. D'autre part, les ressources linguistiques telles que les corpus annotés ou les inventaires de sens (comme [BabelNet](#) ou [WordNet](#)) sont essentiels pour le développement et l'évaluation des modèles.

Les applications de la désambiguïsation en sens sont diverses et touchent de nombreux domaines. Dans la traduction automatique, par exemple, choisir le sens correct d'un mot dans la langue source peut grandement améliorer la qualité de la traduction en évitant les erreurs de sens ([Vickrey et al., 2005](#)). De même, dans les moteurs de recherche, la désambiguïsation en sens peut aider à mieux comprendre les intentions de l'utilisateur en identifiant précisément les mots clés dans la requête de recherche ([Stokoe et al., 2003](#)). Dans le domaine de l'analyse des sentiments, savoir si un mot est utilisé de manière positive ou négative peut influencer l'interprétation globale d'un texte ([Sumanth and Inkpen, 2015](#)).

Cependant, malgré les progrès réalisés dans le domaine, la désambiguïsation en sens reste un défi ouvert en raison de la complexité et la subtilité du langage naturel. De nombreuses ambiguïtés sont difficiles à résoudre même pour les humains, ce qui rend la

tâche d'autant plus difficile pour les systèmes automatiques. De plus, la variabilité du langage dans différents contextes et domaines complique la tâche de créer des systèmes de désambiguïsation en sens généralisables et robustes.

Malgré ces défis, les avancées récentes dans le domaine de l'apprentissage, en particulier avec l'utilisation de réseaux neuronaux profonds, ont permis d'obtenir des performances prometteuses dans la désambiguïsation en sens ([Navigli, 2009](#)).

1.2 État de l'art

La compréhension et le TAL en français ont connu des avancées significatives. D'une part grâce au développement de modèles permettant d'effectuer la tâche de désambiguïsation lexicale comme le modèle FlauBERT. D'autre part, à travers l'élaboration de corpus de qualité avec des annotations manuelles, dites annotations *golds* permettant d'améliorer et d'entraîner des modèles déjà existants tels que le projet FrSemCor. Ces initiatives fournissent des ressources importantes et des modèles spécialisés pour l'annotation sémantique et l'apprentissage de modèles de langue contextualisés, respectivement. Cet état de l'art se penchera sur ces deux types de contributions majeures, leurs méthodologies, leurs résultats et leurs impacts dans le domaine du TAL. Nous nous intéresserons particulièrement au projet permettant d'améliorer les ressources car cela correspond à l'objectif final de notre recherche.

1.2.1 FlauBERT : un modèle de langue évalué sur la tâche de désambiguïsation lexicale

L'avènement de FlauBERT représente une avancée pour le TAL en français, en proposant des modèles de langue contextualisés pré-entraînés adaptés à cette langue ([Le et al., 2020](#)). Contrairement à la plupart des modèles de langue contextualisés qui étaient traditionnellement entraînés sur des corpus anglais, FlauBERT répond à ce besoin en offrant des modèles conçus spécifiquement pour le français, accompagnés d'un cadre d'évaluation complet, FLUE (*French Language Understanding Evaluation*).

Les données d'apprentissage de FlauBERT proviennent d'un vaste corpus de 71 Go, incluant des textes provenant de Wikipedia, de livres et du *Common Crawl*. Ces données sont essentielles pour la robustesse et la généralisation des modèles. FlauBERT est entraîné en utilisant l'objectif de modèle de langue masquée (MLM), une approche qui permet au modèle de générer du texte en se basant sur le contexte environnant.

FlauBERT se décline en deux principales variantes :

- FlauBERT_{BASE} : version plus légère pour des applications moins exigeantes.
- FlauBERT_{LARGE} : version plus complexe pour des performances accrues.

FlauBERT a été évalué sur diverses tâches de TAL, il a été observé qu'en comparaison avec les modèles multilingues comme par exemple mBERT et autres modèles monolingues français, FlauBERT démontre des avantages significatifs :

- Des performances supérieures dans des tâches telles que l'analyse syntaxique et la désambiguïsation lexicale.
- Une adaptabilité, avec une capacité à capturer des nuances linguistiques spécifiques au français, améliorant ainsi la qualité des résultats.
- Une flexibilité, notamment à travers les deux variantes (BASE et LARGE) permettent une application adaptée selon les besoins spécifiques des tâches.

Ces variantes permettent une flexibilité d'application selon les besoins spécifiques des tâches linguistiques.

Les évaluations de FlauBERT sur diverses tâches de TAL, telles que la classification de texte, l'analyse syntaxique et la désambiguïsation lexicale, ont démontré des performances souvent supérieures à celles des modèles multilingues comme mBERT. Contrairement aux modèles multilingues et autres modèles monolingues français, FlauBERT excelle particulièrement dans l'analyse syntaxique et la désambiguïsation lexicale, grâce à sa capacité à capturer les nuances linguistiques spécifiques au français. (cf. la figure 1.1)

Tâche Section Mesure	Classification			Paraphrase Acc.	NLI Acc.	Constituants		Dépendances		Désambiguïsation	
	Livres Acc.	DVD Acc.	Musique Acc.			F ₁	POS	UAS	LAS	Noms F ₁	Verbes F ₁
État de l'art ant.	91.25 ^c	89.55 ^c	93.40 ^c	66.2 ^d	80.1/ 85.2^e	87.4 ^a		89.19 ^b	85.86 ^b	-	43.0 ^h
Sans pré-entr.	-	-	-			83.9	97.5	88.92	85.11	50.0	-
FastText	-	-	-			83.6	97.7	86.32	82.04	49.4	34.9
mBERT	86.15 ^c	86.9 ^c	86.65 ^c	89.3 ^d	76.9 ^f	87.5	98.1	89.5	85.86	56.5	44.9
CamemBERT	93.40	92.70	94.15	89.8	81.2	88.4	98.2	91.37	88.13	56.1	51.1
FlauBERT _{BASE}	93.40	92.50	94.30	89.9	81.3	89.1	98.1	91.56	88.35	54.9/ 57.9^g	47.4

FIGURE 1.1 – Tableau des performances finales sur les tâches de FLUE (Kitaev et al., 2019). (Constant et al., 2013). (Eisenschlos et al., 2019). (Chen et al., 2017). (Martin et al., 2020). (Segonne et al., 2019) (Tableau 2 de (Le et al., 2020))

Cependant, une critique récurrente concerne la qualité des annotations dans les corpus utilisés pour l'entraînement de FlauBERT, qui pourraient être améliorées pour obtenir des résultats encore plus précis et fiables. L'article de Le et al. (2020) suggère que l'amélioration continue de la qualité des annotations pourrait ainsi bénéficier à la précision et à la fiabilité des résultats de FlauBERT dans des contextes linguistiques diversifiés.

1.2.2 Wiktionsnaire : un inventaire de sens pour les annotations en sens du français

L'utilisation de Wiktionsnaire comme ressource pour la désambiguïsation lexicale, en particulier pour les verbes en français, a été explorée par Delli Bovi et al. (2017). Ils ont évalué l'utilisation d'Eurosense, un corpus multilingue annoté automatiquement avec des sens BabelNet, comme données d'entraînement pour la WSD en français. Cependant, leur évaluation a révélé que la qualité des annotations n'était pas suffisante pour une désambiguïsation en sens supervisée des verbes français. Ils ont donc proposé d'utiliser Wiktionsnaire, un dictionnaire en ligne multilingue édité collaborativement, comme nouvelle ressource pour la désambiguisation en sens. Wiktionsnaire fournit à la fois un

inventaire de sens et des exemples étiquetés manuellement, pouvant être utilisés pour former des systèmes de WSD supervisés et semi-supervisés.

[Segonne et al. \(2019\)](#) ont créé FrenchSemEval, un nouveau jeu de données d'évaluation pour la désambiguïsation des verbes français, annoté manuellement avec des sens de Wiktionnaire. Ils ont comparé statistiquement le corpus d'exemples de Wiktionnaire avec SemCor, un corpus de référence pour l'anglais. Les premières expériences de WSD utilisant Wiktionnaire comme base ont montré des résultats encourageants, mais ont également mis en évidence le gain de performance potentiel qui pourrait être obtenu en augmentant le nombre d'exemples d'entraînement (cf. la figure 1.2).

Language	Corpus (# annotations)	AMBIG_trainSI		AMBIG_fullSI	
		type	token	type	token
English	SemCor (88334)	1.97	7.91	3.24	10.94
	SenseEval2 (517)	4.90	6.7	7.58	10.28
	SemEval 2007 (296)	5.15	6.89	7.78	10.17
	SenseEval 2015 (251)	5.69	6.25	8.48	9.16
French	Wiktionary (55206)	1.66	5.49	1.74	5.68
	FSE (3199)	6.02	6.74	6.15	6.91

FIGURE 1.2 – Taux d'ambiguïté des verbes, en anglais usuel ensembles, ainsi que dans l'ensemble de formation (Wiktionnaire) et l'ensemble d'évaluation (FSE) français. **AMBIG_trainSI** correspond à utiliser pour le nombre de sens l'inventaire des sens dans le corpus d'entraînement correspondant, tandis qu'**AMBIG_fullSI** correspond à utiliser l'inventaire des sens complet. (Tableau 3 de l'article de ([Segonne et al., 2019](#)))

L'étude suggère que l'inventaire de sens de Wiktionnaire est approprié pour une annotation de qualité des corpus, et que l'utilisation des exemples de Wiktionnaire comme données d'entraînement pour la WSD des verbes français offre des résultats prometteurs. On note que l'ajout d'un nombre modéré d'exemples supplémentaires puisse améliorer la désambiguïsation. [Segonne et al. \(2019\)](#) envisagent même l'automatisation de la sélection et de l'annotation d'instances supplémentaires comme une voie possible pour améliorer la désambiguïsation des sens des verbes.

1.2.3 SemEval2013 : un corpus textuel d'annotation en sens de qualité pour la désambiguïsation lexicale en français

SemEval2013, ou *Semantic Evaluation*, est une série de compétitions internationales visant à évaluer les systèmes de TAL sur diverses tâches sémantiques.

L'édition 2013 de la compétition SemEval ([Dzikovska et al., 2013](#)), marque la septième occurrence de cet événement. Elle a réuni des équipes du monde entier, cherchant à mesurer et comparer leurs approches en matière de compréhension et d'analyse du langage naturel. Chaque tâche de SemEval2013 a été conçue pour tester des aspects spécifiques

de la sémantique linguistique.

Parmi les tâches principales de SemEval2013, la simplification lexicale en anglais (Spezia et al., 2012) visait à remplacer des mots complexes dans une phrase par des synonymes plus simples sans altérer le sens global. Une autre tâche notable était l'analyse des sentiments sur Twitter (Nakov et al., 2013), où les systèmes devaient classifier les tweets en sentiments positifs, négatifs ou neutres. De plus, la tâche de similarité sémantique textuelle (Agirre et al., 2013) mesurait dans quelle mesure deux phrases étaient similaires en termes de sens, sans oublier la tâche de désambiguïsation lexicale (Navigli et al., 2013). Ces défis reflétaient une diversité d'applications pratiques du TAL, mettant en lumière les défis techniques et linguistiques que rencontrent les chercheurs dans ce domaine.

Les systèmes participants à SemEval2013 étaient évalués selon des critères stricts, spécifiques à chaque tâche. Les métriques d'évaluation comprenaient généralement la précision, le rappel, et la F-mesure, entre autres. Chaque équipe soumettait ses résultats pour une ou plusieurs tâches, et ces résultats étaient ensuite comparés à des *baselines*¹ et à ceux des autres participants. Cette évaluation permettait de distinguer les approches les plus efficaces et innovantes, stimulant ainsi le progrès dans le domaine du TAL.

L'impact de SemEval2013 sur la recherche en traitement automatique des langues a été conséquent. En fournissant un cadre pour l'évaluation comparative, SemEval2013 a encouragé l'innovation et l'amélioration continue des techniques de TAL. Les données et résultats obtenus lors de cette édition sont devenus des références clés pour les chercheurs et les professionnels du secteur. De plus, les avancées réalisées grâce à SemEval2013 ont trouvé des applications pratiques dans divers domaines, notamment la traduction automatique, l'extraction d'informations, et l'analyse des médias sociaux.

Dans le cadre de notre mémoire, nous allons nous intéresser à la tâche 12 du SemEval2013 qui vise plus particulièrement le défi concernant la désambiguïsation lexicale.

1. Une « *baseline* » dans le contexte de l'évaluation des systèmes de TAL se réfère à une méthode de référence simple et souvent rudimentaire, utilisée comme point de comparaison pour évaluer la performance des systèmes plus sophistiqués ou des modèles développés.

Chapitre 1. Définitions et État de l'art

Chapitre 2

Corpus et Outils de recherche

Ce chapitre expose les choix méthodologiques effectués, en commençant par une présentation des corpus utilisés, comme celui de SemEval2013, et des ressources lexicographiques, telles que WordNet et BabelNet, qui ont servi d'inventaires de sens pour l'annotation. Nous décrivons également la constitution et l'annotation du corpus FrSemCor en français, mettant en lumière l'avantage des « *supersenses* » pour faciliter une analyse sémantique affinée.

Ces différentes ressources et outils nous ont permis d'utiliser et de tester un modèle pour la désambiguïsation lexicale dans un contexte, en évaluant sa performance sur divers ensembles de données.

2.1 Corpus

2.1.1 Description du corpus SemEval2013

La tâche 12 de SemEval2013 (Navigli et al., 2013) s'est concentré sur la désambiguïsation sémantique des mots (WSD). Historiquement, la désambiguïsation de sens a évolué à partir d'inventaires de sens correspondant à des références standardisées comme WordNet. WordNet est une base de données lexicale qui organise les mots en synonymes et en ensembles de sens distincts, fournissant une référence structurée pour la désambiguïsation. Plus récemment, des initiatives telles que SemEval2013 ont introduit des inventaires de sens plus étendus et multilingues comme BabelNet, qui combine les données de WordNet avec des informations de Wikipédia pour une couverture sémantique plus riche et diversifiée.

Dans le cadre de la tâche 12, les systèmes participants ont été invités à annoter les noms dans le corpus de test avec le sens le plus approprié à partir de l'inventaire de sens de BabelNet, de WordNet ou de Wikipedia. BabelNet 1.1.1 a été utilisé comme l'inventaire principal pour cette tâche, offrant une large couverture multilingue grâce à sa combinaison de données provenant de WordNet et de Wikipédia. Les performances des systèmes ont été évaluées en termes de précision et de rappel par rapport à une référence appelée « *Most Frequent Sense* » (MFS) (Arora et al., 2016), qui utilise le sens le plus fréquent défini par WordNet comme *baseline*.

Le corpus de la tâche 12 comprend un ensemble de 13 articles issus des éditions 2010, 2011 et 2012 de l'atelier de traduction automatique statistique (WSMT).

Chaque article est disponible dans cinq langues : l'anglais, le français, l'allemand, l'espagnol et l'italien.

Pour l'italien, les articles ont été traduits manuellement de l'anglais par des locuteurs na-

Chapitre 2. Corpus et Outils de recherche

tifs italiens avec le soutien d'un conseiller ayant pour langue maternelle l'anglais, assurant ainsi la qualité de la traduction et une meilleure cohérence linguistique.

Le corpus français a été annoté manuellement pour un concours informatique dans lequel les participants devaient utiliser ce corpus comme jeu de données de validation afin d'évaluer la précision des performances de leur modèle d'annotation automatique en sens. La figure 2.1 présente le format du corpus SemEval2013, c'est un fichier XML, chaque mot a un lemme, une étiquette de partie du discours et si c'est un nom commun, une annotation en sens (encadré en blanc sur la figure).

```
<sentence id="d001.s001" >
  <word surface_form="Le" lemma="le" pos="DET:ART" />
  <word surface_form="groupe" lemma="groupe" pos="NOM" wn30_key="group%1:03:00::;grouping%1:03:00::" id="d001.s001.t001" />
  <word surface_form="des" lemma="du" pos="PRP:det" />
  <word surface_form="Nations_Unies" lemma="nations_unies" pos="NE" wn30_key="un%1:14:00::;united_nations%1:14:00::" id="d001.s001.t002" />
```

FIGURE 2.1 – Exemple d'annotation manuelle dans le corpus SemEval2013-fr.

Ce corpus contient 306 phrases avec un total de 9 648 *tokens*, tous annotés avec une étiquette de partie du discours.

Les noms communs représentent la proportion la plus élevée, bien que le corpus inclut également des verbes, des adjectifs ainsi que d'autres catégories grammaticales. La figure 2.2 représente les proportions de nos données concernant le corpus SemEval2013.

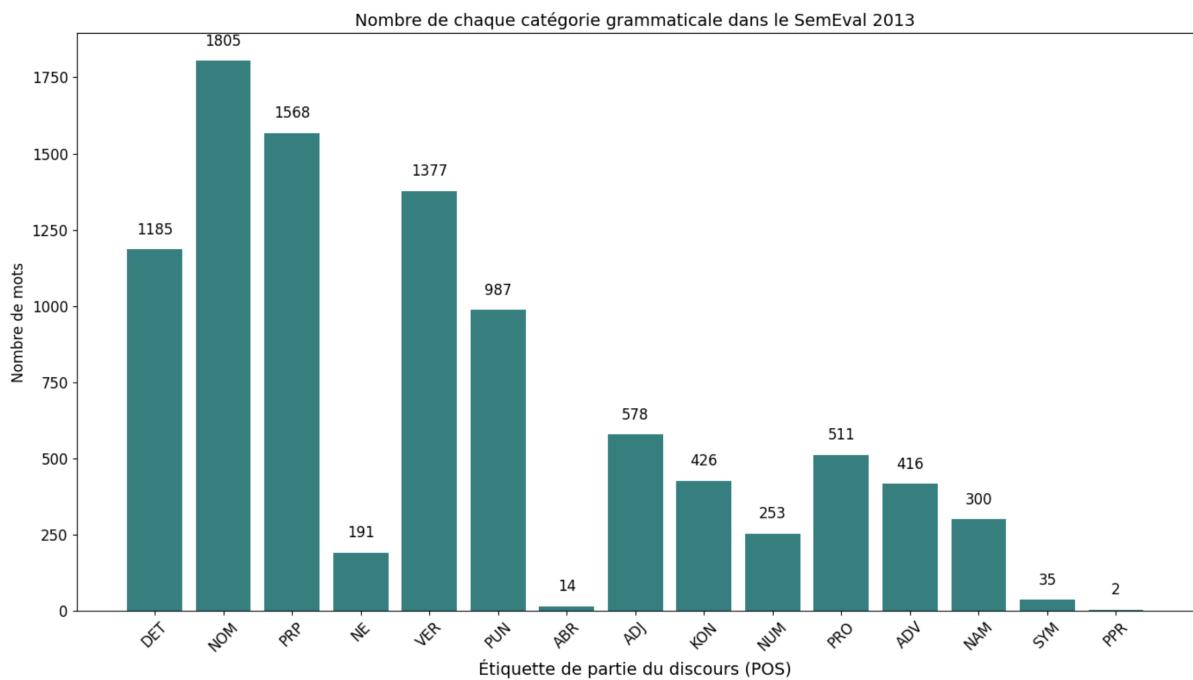


FIGURE 2.2 – Répartition des étiquettes de partie du discours dans SemEval2013.

Parmi les 1 445 *tokens* annotés en sens dans ce corpus, il existe trois types d'étiquettes de partie de discours possibles :

- NOM : noms communs (1 367)
- NAM : noms propres (3)

- NE : entités nommées (75)

Dans les noms non-annotés en sens nous retrouvons des mots tels que :

mots_sans_wn30_key
Mots sans wn30_key
projets
affectations
feux_de_la_rampe
raison
publication
plupart
avancée
pays
riches
pauvres
réchauffement

FIGURE 2.3 – Extraits des noms ne possédant pas d’annotation en sens

Nous considérons seulement les *tokens* ayant comme étiquette « NOM », cependant, certaines de ces instances nominales ne possèdent pas d’annotations en sens, soit en raison de l’absence de correspondance adéquate dans les inventaires de sens, soit pour d’autres raisons possibles telles qu’une ambiguïté trop élevée, un manque de contexte ou encore une complexité syntaxique ([Navigli, 2009](#)).

De plus, on remarque qu’un peu plus de 25 % des noms communs n’ont pas reçu d’annotation en sens. En effet, sur 1 805 noms communs, 1 367 possèdent une annotation en sens et 438 n’en ont pas. La figure [2.4](#) représente les proportions des noms possédant une annotation en sens et ceux qui n’en possèdent pas.

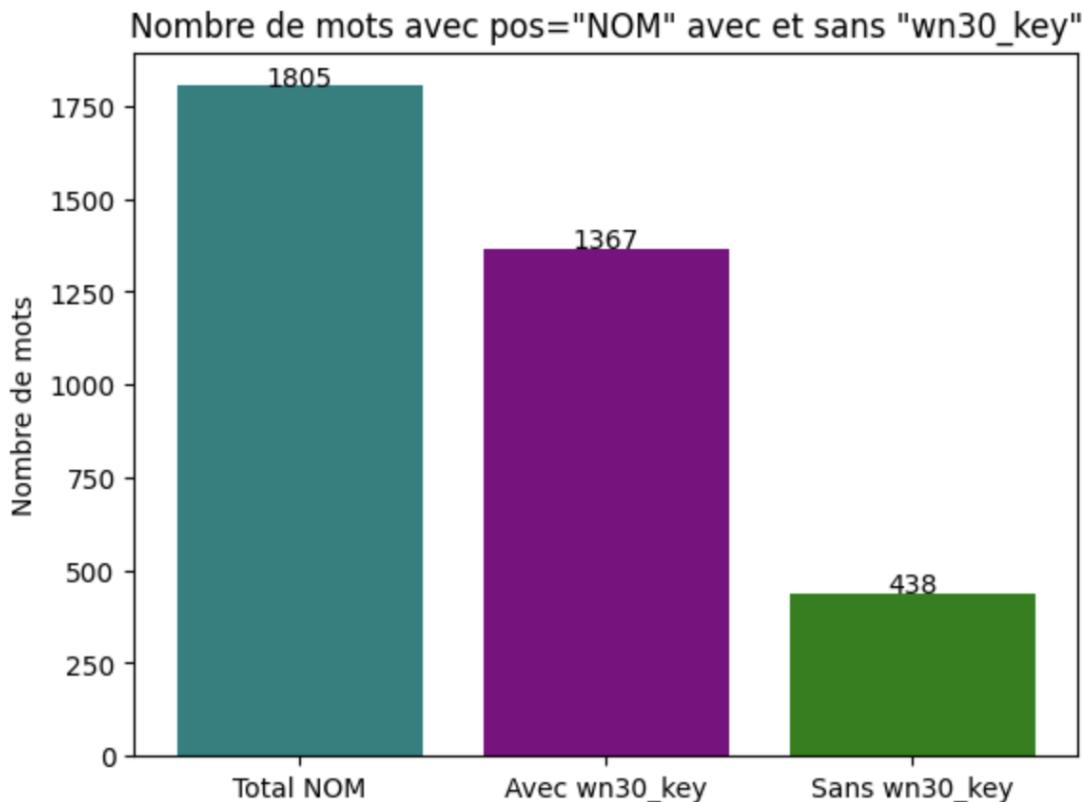


FIGURE 2.4 – Diagramme représentant les proportions des noms possédant une annotation en sens et ceux qui n'en possèdent pas

Trois équipes ont soumis un total de sept systèmes pour la tâche, chacun utilisant des approches variées basées sur des graphes de sens de WordNet ou BabelNet. Les résultats ont montré que les systèmes utilisant BabelNet ont généralement surpassé la *baseline* MFS, démontrant ainsi l'efficacité de l'utilisation de BabelNet pour la désambiguïsation de sens multilingue. Les performances variaient cependant selon les langues et les types d'entités traitées (comme les expressions polylexicales ou les entités nommées), mettant en lumière les défis persistants dans ce domaine passionnant de la recherche en traitement automatique du langage naturel.

2.1.2 Description du corpus FrSemCor : un corpus français annoté avec des *supersenses*

FrSemCor ([Barque et al., 2020](#)) est un projet basé sur la construction d'un corpus annoté en français, dans lequel les **noms** sont manuellement étiquetés avec des classes sémantiques, appelées *supersenses*. L'objectif principal de ce projet est de pallier le manque de données annotées sémantiquement.

Cette ressource vise à combler un besoin de ressources linguistiques riches capable de capturer les nuances des sens du français ainsi que fournir une référence de qualité, basée sur une annotation sémantique rigoureuse supervisée par des experts en sémantique lexicale. Le corpus annoté comprend plus de 12 000 occurrences de noms communs dans un corpus d'arbres syntaxiques en français.

Les étiquettes utilisées dans les corpus annotés sémantiquement se divisent en deux catégories principales : celles liées à des ensembles de sens prédéfinis, comme dans un dictionnaire, et celles correspondant à des classes sémantiques générales non pré-associées à un lexique spécifique. Par exemple, dans le corpus anglais SemCor, les occurrences d'un mot sont associées à une liste prédéfinie de sens (représentés par des *synsets*) provenant de WordNet. Ce dernier, connu pour ses distinctions lexicales, est organisé hiérarchiquement en différents niveaux sémantiques ; les « *Unique Beginners* »¹ (UBs) constituant le niveau le plus élevé de l'ontologie de WordNet.

Le jeu d'étiquettes sémantiques utilisé pour annoter les noms dans le corpus français est dérivé des WordNet UBs, qui n'étaient pas initialement conçus pour l'annotation mais pour faciliter le travail lexicographique en regroupant des *synsets* partageant le même hyperonyme général. Pour intégrer les UBs dans un jeu d'étiquettes opérationnel, plusieurs ajustements ont été nécessaires pour clarifier les définitions et introduire des opérateurs générant des *tags*² sémantiques complexes pour tenir compte des phénomènes contextuels ou lexicaux.

La méthodologie adoptée pour FrSemCor se distingue par un équilibre entre l'annotation manuelle et l'utilisation d'outils computationnels. Les étapes principales incluent :

- Sélection du corpus : Choix des textes représentatifs de la langue française contemporaine.
- Directives d'annotation : Élaboration de règles précises pour guider les annotateurs.
- Formation des annotateurs : Assurance de la cohérence et de la fiabilité des annotations.
- Ressources lexicales : Utilisation de bases de données existantes et de connaissances d'experts pour affiner les annotations.

FrSemCor explore donc les modèles de distribution des *supersenses* en français, offrant des analyses statistiques détaillées qui révèlent les structures sémantiques sous-jacentes. Les résultats montrent la fréquence et la distribution des *supersenses* selon différentes classes de mots et contextes syntaxiques, fournissant ainsi des « *insights* » intéressants pour la désambiguïsation de sens des mots et l'analyse syntaxique.

Les annotations sémantiques de FrSemCor ont des applications potentielles variées allant de l'amélioration des systèmes de traduction automatique à l'optimisation des algorithmes de recherche d'informations. La disponibilité de FrSemCor comme ressource ouverte encourage également la collaboration et l'innovation dans la recherche en sémantique française.

1. Les « *Unique Beginners* » dans WordNet sont les concepts les plus généraux au sommet de la hiérarchie sémantique. Ils servent de points de départ pour organiser les *synsets*. Exemples d'UBs incluent « entity » (entité) et « event » (événement), représentant des catégories sémantiques très larges et abstraites.

2. Un *tag* sémantique est une étiquette ou un marqueur utilisé pour annoter des éléments de texte afin de refléter leur signification ou leur rôle dans un contexte donné.

Qu'est-ce qu'un *supersense* ?

Les *supersenses* (Flekova and Gurevych, 2016), également appelés fichiers lexicographiques ou champs sémantiques, sont des étiquettes sémantiques de haut niveau utilisées pour catégoriser les mots en fonction de leurs significations générales. Initialement créés pour organiser les ressources lexicales sémantiques comme WordNet, les *supersenses* permettent de regrouper les mots en catégories plus larges et moins détaillées que les sens individuels de WordNet. Cela permet de simplifier certaines tâches de traitement automatique du langage naturel.

Par exemple, WordNet peut distinguer des nuances très fines entre les sens des mots, ce qui peut se révéler compliqué pour des applications pratiques comme la traduction automatique ou la récupération d'informations. Les *supersenses*, en revanche, offrent des étiquettes plus grossières et gérables. Il existe 24 étiquettes pour les noms (cf le tableau 2.1).

FrSemCor	
Supersenses	Generalisation
Act, Event, Process, Phenomenon	Dynamic _situation
Attribute, State, Feeling, Relation	Stative _situation
Animal, Person	Animate _entity
Body, Object, Plant	Natural _Object
Cognition, Communication	Informational _object
Part, Quantity, Group Artifact, Food, Institution, Location, Motive, Possession, Shape, Substance, Time, Tops	Quantification Other

TABLEAU 2.1 – *Supersenses* utilisés en annotation dans FrSemCor (à gauche), regroupés en classes plus générales (à droite). Les *supersenses* barrés sont les UBs qui n'ont pas été retenus, et les *supersenses* en gras sont ceux qui ont été ajoutés (cf. Tableau 1 de (Barque et al., 2020))

Les *supersenses* sont utiles pour diverses tâches de TAL telles que le *parsing* de dépendances³, la reconnaissance des entités nommées, la réponse aux questions non factuelles, la génération de questions, l'étiquetage des rôles sémantiques, le profilage de la personnalité, la similarité sémantique et la détection de métaphores. En résumé, les *supersenses* permettent de simplifier et d'améliorer la gestion sémantique des mots dans les applications de traitement du langage.

3. Le *parsing* de dépendance (ou analyse de dépendance) est une méthode en TAL qui consiste à analyser la structure grammaticale d'une phrase en identifiant les relations entre les mots. Contrairement à l'analyse syntaxique en constituants, qui se concentre sur les groupes de mots (comme les syntagmes nominaux ou verbaux), l'analyse de dépendance met en évidence les relations hiérarchiques directes entre les mots sous forme de dépendances, où chaque mot dépend d'un autre mot appelé « tête ». Cela permet de représenter la structure d'une phrase sous forme d'un graphe orienté, facilitant la compréhension des relations syntaxiques entre les mots.

2.1.3 L'annotation en *Supersenses*

L'annotation en *supersenses* est une méthode dite « à gros grain », qui consiste à regrouper les différents sens d'un mot sous une seule catégorie sémantique générale. Cette approche simplifie le processus d'annotation sémantique ([Barque et al., 2020](#)).

Comme mentionné précédemment, la tâche de désambiguïsation lexicale (WSD) est complexe. En effet, les résultats montrent qu'une annotation « à gros grain » offre de meilleures performances qu'une annotation « à fin grain » (Accord inter-annotateur avec un gros grain : 94 ([Navigli et al., 2007](#)) contre accord inter-annotateur avec WordNet : 70 ([Snyder and Palmer, 2004](#)).) Cela peut s'expliquer par le fait que les annotations sont moins nombreuses, facilitant donc l'apprentissage des modèles utilisés.

Il existe plusieurs méthodes pour réaliser cette annotation : soit en utilisant un inventaire de classes sémantiques générales, soit en utilisant un lexique avec une liste des sens des mots ([Ciaramita and Johnson, 2003](#); [Schneider et al., 2012](#); [Pedersen et al., 2016](#)), soit en réduisant la polysémie lorsque le lexique possède une liste des sens des mots ([Navigli, 2006](#); [PALMER et al., 2007](#)).

Dans WordNet, les *supersenses* sont appelés « *Unique Beginners* » (section [2.1.2](#)). À partir de cet inventaire de sens, des étiquettes sémantiques ont été créées pour des travaux lexicographiques ([Miller et al., 1990](#); [Fellbaum, 1998](#)) et peuvent être aujourd'hui utilisées pour les *supersenses*.

Les étiquettes en anglais sont les suivantes : *Act, Animal, Artifact, Attribute, Body, Cognition, Communication, Event, Feeling, Food, Group, Location, Motive, Object, Person, Phenomenon, Plant, Possession, Process, Quantity, Relation, Shape, State, Substance, Time, Tops*.

Par la suite, ces étiquettes ont été adaptées et modifiées pour le corpus FrSemCor ([Barque et al., 2020](#)).

Pour le français, plusieurs corpus annotés en *supersenses* sont disponibles, offrant des annotations de référence (*gold*).

Le corpus FrSemCor est l'un des principaux corpus annotés en sens, utilisant l'inventaire de sens de WordNet pour le français. D'autre part, le corpus Sequoia intègre des annotations provenant de plusieurs sources, telles qu'Europarl, l'Est Républicain, Wikipédia-fr et l'EMEA (*Europe, Middle East & Africa*) ([Candito and Seddah, 2012](#)). Enfin, le corpus EuroSens a été annoté automatiquement en utilisant les inventaires de sens de BalbelNet et d'Europarl ([Delli Bovi et al., 2017](#)).

Ces ressources fournissent des données intéressantes pour la recherche en désambiguïsation lexicale et en traitement automatique des langues pour le français.

2.2 Ressources lexicographiques

2.2.1 WordNet : un inventaire de sens pour l'anglais

WordNet ([Miller, 1994](#)) est une base de données lexicale en ligne conçue pour faciliter le traitement du langage naturel en organisant les mots de l'anglais en ensembles de synonymes appelés *synsets*. Chaque *synsets* représente un concept lexical distinct, regroupant des substantifs, des verbes, des adjectifs et des adverbes en fonction de leur significations communes. Par exemple, des mots tels que « arbre » et « plante » seraient liés hiérarchiquement en tant que hyperonymes et hyponymes respectivement, décrivant des relations de généralisation et de spécifications entre les concepts.

La structure de WordNet permet également de représenter les relations sémantiques entre les mots. Outre les synonymes, qui continuent la relation de base, WordNet inclut des relations comme les antonymes, qui expriment des sens opposés (par exemple, « grand » et « petit »), et les relations de métonymie et homonymie, qui décrivent les relations de parties à un tout (par exemple, « roue » est une métonymie de « voiture », et vice-versa) (cf. la figure 2.2)

Relation Sémantique	Catégorie Syntaxique	Exemples
Synonymie (similarité)	N, V, Adj	livre/bouquin, escalader/grimper, triste/malheureux
Antonymie (opposé)	Adj, N, V	sec/mouillé, jour/nuit, monter/descendre
Hyponymie (subordonné)	N	chat (félins), cerise (fruits)
Meronymie (partie)	N	bras/corps, toit/maison, roue/voiture
Troponymie (manière)	V	égorger, décapiter, guillotiner, électrocuter (tuer)

TABLEAU 2.2 – Exemples de relations lexicales dans WordNet.

Cette base de données contient une vaste gamme de formes et de sens de mots avec plus de 118 000 formes de mots différentes et plus de 90 000 sens distincts, totalisant plus de 166 000 paires (forme, sens). Environ 17 % des mots répertoriés dans WordNet sont polysémiques, tandis que 40 % d'entre-eux ont au moins un synonyme enregistré.

WordNet est l'une des ressources de TAL les plus populaires, son but est de répertorier, classifier et mettre en relation de diverses manières le contenu sémantique et lexicale de la langue anglaise.

Le système est disponible sous forme de données électronique téléchargeable localement. Il propose également des interfaces de programmation compatibles avec de nombreux langages dont le langage python que nous utiliserons dans notre projet.

Qu'est-ce qu'un *synset* ?

Dans WordNet, le *synset* (Chaumartin, 2007), ou ensemble de synonymes, constitue l'unité fondamentale sur laquelle repose la base de données. Chaque *synset* regroupe des mots interchangeables qui expriment un sens ou un usage particulier, défini de manière distinctes par ses relations. Par exemple, le *synset* du mot « voiture » : « voiture, auto, automobile » (cf. la figure 2.5).

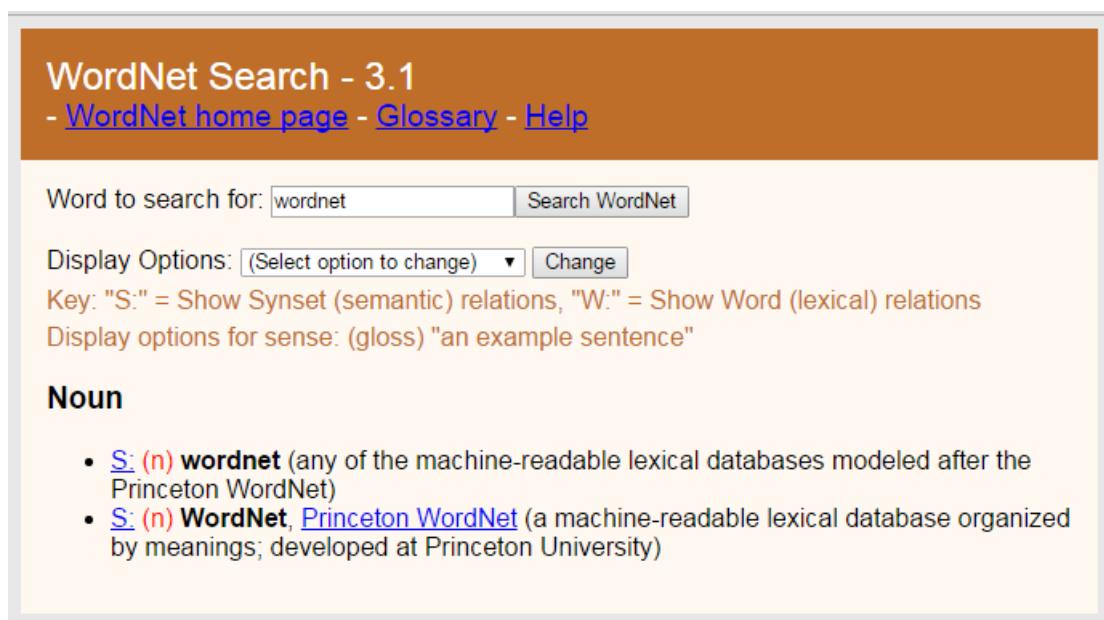


FIGURE 2.5 – Exemple d'affichage des *synsets* dans WordNet

2.2.2 BabelNet : un inventaire de sens multilingue

The screenshot shows the BabelNet interface with a teal header bar. On the left is a menu icon (three horizontal lines). Next to it is the BabelNet logo with the text 'BabelNet v5.3'. Below the header are several input fields and buttons: a search bar containing 'moment', a language dropdown set to 'French', a 'Traduire en...' (Translate to...) dropdown, a search button with a magnifying glass icon, and two buttons for part of speech: '20 Noun' and '0 Verb'. Below these controls, the search term 'moment' is shown in bold with the word 'nom' in parentheses, indicating its part of speech. Three search results are displayed:

- FR moment nom**
L'histoire et la chronologie de l'Univers décrit l'évolution de l'Univers en s'appuyant sur le modèle standard de
heure • histoire de l'Univers • Histoire et chronologie de l'Univers
bn:01188813n Concept ★
An image of a book cover is shown next to the text.
- En théorie des probabilités et en statistique, les moments d'une variable aléatoire réelle sont des indicateurs**
bn:00055597n Concept ★ | probabilités
An image of a graph with multiple bell-shaped curves is shown next to the text.
- Le moment d'un vecteur peut se définir par rapport à un point ou par rapport à un axe orienté.**
moment d'un vecteur
bn:00055596n Concept ★ | English: physics mechanics
An image of a vector space diagram is shown next to the text.

FIGURE 2.6 – Présentation de l’interface de BabelNet

BabelNet est un « dictionnaire encyclopédique » conçu en 2013 par Sapienza NLP Group sous la supervision du professeur Roberto Navigli et de Babelscape (entreprise de deep tech NLP multilingue) (Navigli et al., 2021). En 2023 le dictionnaire prend en compte 600 langues naturelles avec 53 sources, il compte presque 2 milliards de sens, environ 23 millions de *synsets*, 15 millions d’entités nommées et 160 millions de définitions.

BabelNet s’appuie sur diverses ressources telles que WordNet et Wikipedia. En effet, cette plateforme a intégré des notions issues de ces deux sources. Par exemple, elle utilise les *synsets* provenant de WordNet : ceux-ci sont extraits pour constituer la base des notions dans BabelNet, aussi, les relations hyperonymiques et méronymiques sont importées. De plus, les titres des articles de Wikipédia ont été adoptés comme termes principaux, tandis que les redirections ont été considérées comme des synonymes afin d’enrichir les *synsets*. Un alignement entre ces deux ressources a été réalisé pour que les termes similaires puissent se compléter.

La figure 2.6 montre un exemple de l’interface de BabelNet. Par exemple, pour faire des annotations, si on cherche le terme « moment » plusieurs sens apparaissent. Parmi ces sens, on choisit le plus approprié au contexte et on récupère l’identifiant BabelNet (encadré en rouge sur la figure).

2.3 Description du modèle AMuSe-WSD

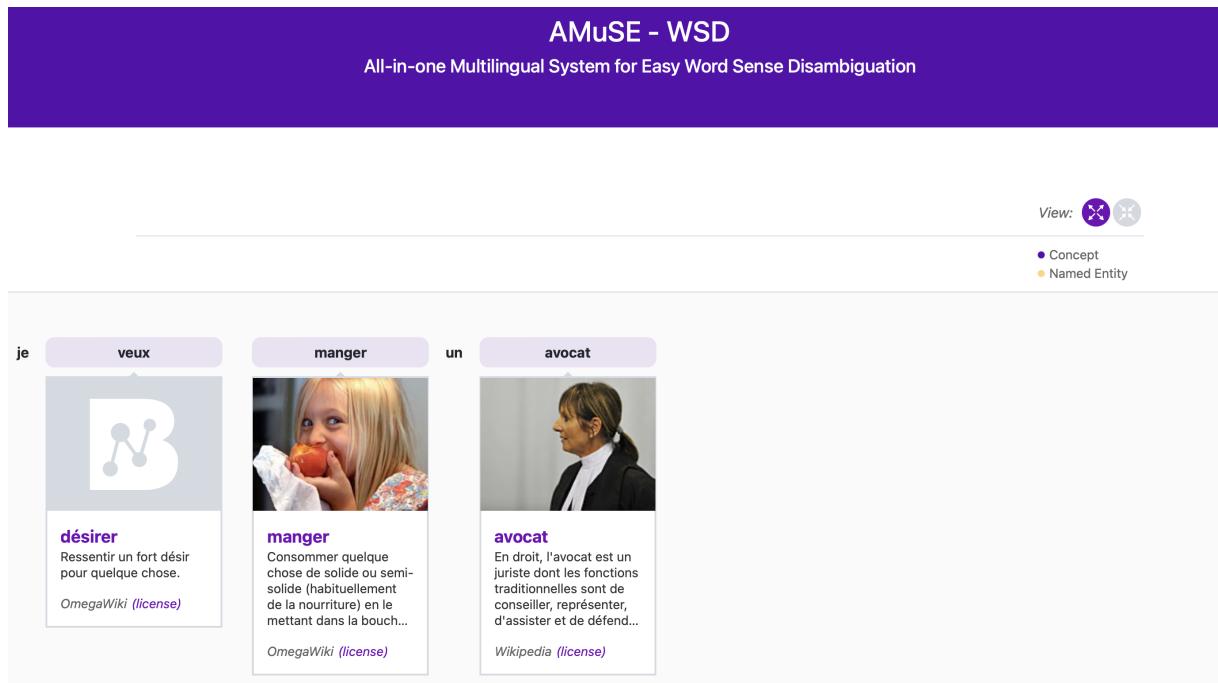


FIGURE 2.7 – Présentation de l’interface du modèle AMuSe-WSD

AMuSe-WSD (Orlando et al., 2021) est un système de désambiguïsation de sens innovant qui se distingue par son approche multilingue et son utilisation d'un modèle neuronal avancé. Contrairement aux approches basées sur les connaissances qui utilisent des lexiques informatiques, et aux approches supervisées traditionnelles, AMuSe-WSD adopte un modèle neuronal de pointe entraîné sur des données annotées en sens.

L’implémentation de systèmes WSD de pointe peut être complexe, nécessitant souvent des modules avancés de prétraitement et de post-traitement pour manipuler le texte de manière précise. Cela peut limiter leur accessibilité aux chercheurs non spécialisés en désambiguïsation en sens. Pour remédier à cela, AMuSe-WSD se présente comme le premier système WSD multilingue tout-en-un, accessible *via* une API REST⁴. Il permet la désambiguïsation en ligne pour une utilisation immédiate ou hors ligne pour minimiser les temps d’inférence.

Il est intéressant de remarquer que ce modèle supporte 40 langues hors-ligne⁵ ainsi que 10 langues en ligne⁶. Cette large couverture linguistique permet de traiter efficacement des textes multilingues, un aspect important pour des applications globales de traitement du langage naturel.

4. Une API REST est une interface de programmation permettant aux applications de communiquer via des requêtes HTTP pour accéder et manipuler des données sur un serveur distant. Elle utilise un ensemble standardisé d’opérations (GET, POST, PUT, DELETE) et échange des données souvent au format JSON ou XML.

5. Pour obtenir la liste complète des 40 langues supportées hors-ligne, vous pouvez visiter [le site officiel d’AMuSe-WSD](#)

6. AMuSe-WSD supporte 10 langues en ligne, telles que l’anglais, le français, l’italien, l’espagnol, l’allemand, le portugais, le néerlandais, le russe, le chinois et l’arabe. ([Orlando et al., 2021](#))

Le *pipeline* d'AMuSe-WSD intègre des outils populaires tels que spaCy et Stanza pour assurer un prétraitement robuste des documents. Cela inclut la segmentation, la tokenisation, la lemmatisation et l'étiquetage en parties du discours (POS tags), pour générer les ensembles de candidats de sens à partir d'inventaires comme WordNet ou BabelNet.

Au cœur du modèle AMuSe-WSD se trouve un système de désambiguïsation des sens basé sur un encodeur⁷. Transformer pré-entraîné pour capturer de manière contextuelle les mots dans le texte, suivie d'une transformation non linéaire pour représenter chaque sens de manière distincte. Il génère ensuite une distribution de scores sur tous les sens possibles dans un inventaire donné.

En termes de performance, AMuSe-WSD a été évalué sur divers *benchmarks* (cf. section 2.3) standards pour la désambiguïsation en sens, montrant des améliorations grâce à l'utilisation de modèles multilingues comme XLM-RoBERTa-large. Les évaluations incluent des jeux de données en anglais ainsi que des évaluations multilingues couvrant des langues comme le français, l'allemand, l'italien et l'espagnol.

Les composants clés d'AMuSe-WSD comprennent des *embeddings* (cf. section 2.3) de sens multilingues qui capturent les significations des mots dans différents contextes linguistiques, un mécanisme d'attention pour se concentrer sur les parties pertinentes du contexte, et un entraînement supervisé sur des corpus annotés comme SemCor, Multilingual Wordnet et BabelNet.

Ces corpus sont annotés manuellement et fournissent une riche source de données pour apprendre les relations entre les mots et leur différents sens dans divers contextes. Les corpus des SemEval, y compris celui du SemEval2013, sont utilisés pour évaluer les performances du modèle. Ces corpus offrent des ensemble de tests annotés manuellement qui permettent de mesurer l'efficacité du modèle sur des phrases réelles en plusieurs langues.

Qu'est-ce qu'un *embedding* ?

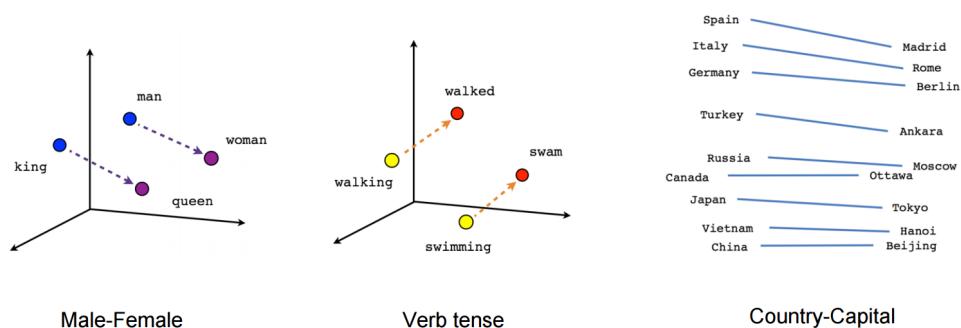


FIGURE 2.8 – Représentation des *Words Embeddings* (Image from polakowo.io)

Les *embeddings* de mots, ou *word embeddings*, sont des représentations numériques qui transforment chaque mot en un vecteur de nombres réels, capturant ainsi des aspects de sa signification et de ses relations avec d'autres mots. Cette transformation repose

7. Un encodeur lit et traite l'intégralité de la séquence de données d'entrée, telle qu'une phrase en anglais, et la transforme en une représentation mathématique compacte

sur l'idée que les mots ayant des contextes similaires ont souvent des significations similaires. Par exemple, dans un modèle de *word embedding* bien entraîné, les mots « chien » et « chat » seront représentés par des vecteurs similaires, reflétant leur proximité sémantique.

Plusieurs méthodes existent pour générer des *embeddings* de mots. Les premières approches, comme le TF-IDF et la réduction de dimensionnalité, ont été remplacées par des techniques plus avancées basées sur des réseaux de neurones. Parmi les plus populaires figurent GloVe, FastText et Word2Vec ([Mikolov et al., 2018](#)), chacune offrant des avantages spécifiques en termes de représentation sémantique et de performance sur diverses tâches de NLP.

Les applications des *words embeddings* sont vastes et variées. Ils sont utilisés pour améliorer la précision des modèles de classification de texte, faciliter la traduction automatique, analyser les sentiments exprimés dans un texte, et même générer du texte de manière cohérente et naturelle. En fournissant des représentations vectorielles de mots qui capturent les nuances de leur signification, les *embeddings* de mots permettent aux algorithmes de NLP de mieux comprendre et manipuler le langage humain.

Un aspect des *embeddings* de mots est leur capacité à être pré-entraînés sur de vastes corpus textuels non annotés, comme des pages *web* ou des corpus de textes littéraires. Ce pré-entraînement permet aux modèles d'apprendre des représentations générales et robustes des mots avant d'être affinés pour des tâches spécifiques. Ce processus de transfert d'apprentissage est essentiel pour adapter les *embeddings* à des domaines particuliers ou à des langues spécifiques, maximisant ainsi leur utilité dans des applications réelles.

Qu'est-ce qu'un *benchmark* ?

En termes de performance, le modèle AMuSe a été évalué sur divers *benchmarks* standard pour la désambiguïsation en sens, démontrant des améliorations grâce à l'utilisation de modèles multilingues comme XLM-RoBERTa-large.

Un *benchmark* est un ensemble de tests standardisés utilisé pour évaluer les performances des modèles de *machine learning*⁸ ou de traitement automatique du langage naturel. Dans le contexte de la désambiguïsation en sens, les *benchmarks* sont des collections de données annotées qui contiennent des phrases avec des mots ambigus, accompagnés des sens corrects de ces mots dans leurs contextes respectifs. Ces ensembles de données permettent de comparer objectivement différents modèles en termes de précision, de rappel et d'autres métriques de performance.

Les *benchmarks* sont importants pour la recherche et le développement en NLP car ils fournissent une base commune pour l'évaluation et la comparaison des modèles. Pour la désambiguïsation en sens, des *benchmarks* comme ceux fournis par les compétitions SemEval (*Semantic Evaluation*) sont largement utilisés. Par exemple, le corpus du SemEval2013 est un *benchmark* bien connu pour évaluer les modèles de désambiguïsation en sens. Ces *benchmarks* contiennent des phrases annotées manuellement avec les sens des mots

8. Le *machine learning*, également connu sous le nom d'apprentissage supervisé, est défini par l'utilisation de jeux de données étiquetés pour entraîner des algorithmes à classer les données ou à prédire les résultats avec précision.

Chapitre 2. Corpus et Outils de recherche

selon des inventaires comme WordNet ou BabelNet, ce qui permet de tester les capacités des modèles à désambiguïser les mots dans des contextes variés et multilingues.

Performance du modèle AMuSe-WSD sur les *benchmarks*

AMuSe-WSD a montré de réelles performances sur ces *benchmarks* standards. Grâce à l'utilisation de modèles multilingues avancés tels que XLM-RoBERTa-large, AMuSe-WSD a pu dépasser les modèles précédents en termes de précision et de rappel ([Orlando et al., 2021](#)).

XLM-RoBERTa-large est un modèle de type *Transformer*⁹ pré-entraîné sur un large corpus multilingue, ce qui lui permet de capturer des représentations contextuelles riches et de mieux gérer la désambiguïsation des sens dans différentes langues.

Modèle	English datasets						Multilingual datasets		
	SE2	SE3	SE07	SE13	SE15	ALL	SE13	SE15	XL-XSD
WSD Modules									
BERT-large	76.3	73.2	66.2	71.7	74.1	73.5	-	-	-
(Conia and Navigli, 2020)	77.1	76.4	70.3	76.2	77.2	76.4	-	-	-
(Scarlino et al., 2020)	78.0	77.1	71.0	77.3	83.2	77.9	78.3	70.8	-
(Blevins and Zettlemoyer, 2020)	79.4	77.4	74.5	79.7	81.7	79.0	-	-	-
(Bevilacqua and Navigli, 2020)	80.8	79.0	75.2	80.7	81.8	80.1	80.3	70.7	-
(Conia and Navigli, 2021)	80.4	77.8	76.2	81.8	83.3	80.2	-	-	-
End-to-End Systems									
(Moro et al., 2014)	67.0	63.5	51.6	66.4	70.3	65.5	65.6	-	52.9
(Papandrea et al., 2017)	73.8	70.8	64.2	67.2	71.5	-	-	-	-
(Scozzafava et al., 2020)	71.6	72.0	59.3	72.2	75.8	71.7	73.2	66.2	57.7
AMuSe-WSD BERT-large	80.6	78.4	76.5	81.0	82.7	80.2	-	-	-
AMuSe-WSD XLMR-large	79.5	77.6	74.1	79.9	83.4	79.3	80.0	73.0	67.3
AMuSe-WSD XLMR-base	77.8	76.0	72.1	77.7	81.5	77.5	76.8	73.0	66.2
AMuSe-WSD m-MiniLM	76.3	72.4	69.5	76.1	77.8	75.1	74.5	69.6	63.9

TABLEAU 2.3 – Résultats de la désambiguïsation lexicale en anglais avec les scores F1 sur Senseval-2 (SE2), Senseval-3 (SE3), SemEval-2007 (SE07), SemEval-2013 (SE13), SemEval-2015 (SE15), ainsi que la concaténation de tous les ensembles de données (ALL). Nous incluons également les résultats sur les ensembles de données multilingues SemEval-2013 (SE13) et SemEval-2015 (SE15). (cf. Tableau 1 de l'article d'[Orlando et al. \(2021\)](#))

9. Un *Transformer* est un modèle de réseau de neurones introduit en 2017 par ([Liu et al., 2023](#)), utilisé principalement en traitement du langage naturel. Il se distingue par son mécanisme d'attention, qui permet de traiter des relations complexes dans les données séquentielles, améliorant ainsi la compréhension contextuelle.

Chapitre 3

Évaluation du modèle AMuSe-WSD sur le corpus SemEval2013

L'objectif de la première partie de notre travail, est d'évaluer les performances du modèle AMuSe-WSD avec les données du corpus SemEval2013 et d'en étudier les résultats.

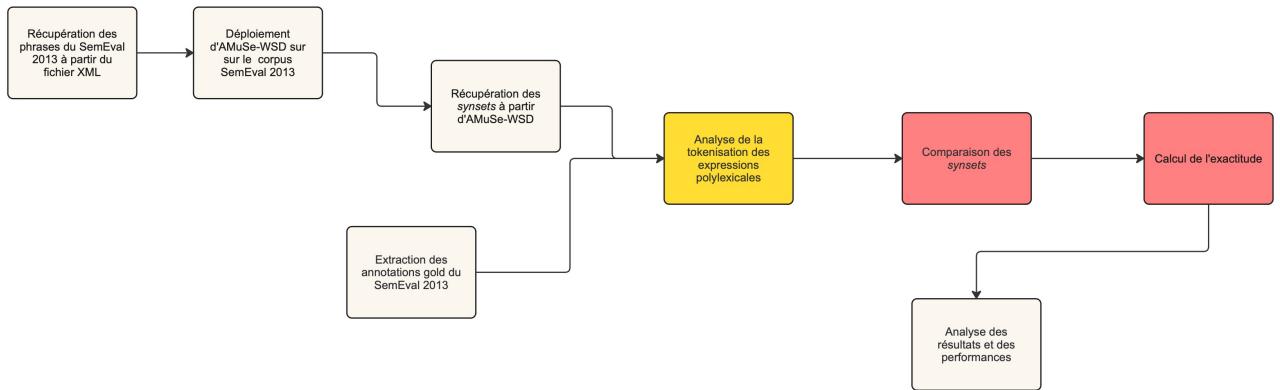


FIGURE 3.1 – Schéma récapitulatif de notre utilisation du modèle AMuSe-WSD sur le corpus SemEval2013

3.1 Récupération des données du corpus SemEval2013 obtenu par le modèle AMuSe-WSD

Dans cette étude, nous avons réalisé une évaluation de la performance du système AMuSe-WSD en comparant les annotations *golds* du corpus SemEval2013 avec les annotations générées par AMuSe-WSD.

Pour évaluer AMuSe-WSD sur le corpus SemEval2013, nous avons d'abord converti le corpus du format XML au format JSON. Cette conversion a été nécessaire car AMuSe-WSD accepte les données au format JSON, qui est plus facilement manipulable par les API.

Le format XML, bien que largement utilisé pour l'annotation linguistique, nécessite souvent un pré traitement supplémentaire pour en extraire les informations pertinentes telles que les balises, les structures hiérarchiques, etc.

En passant au format JSON, nous avons simplifié l'accès aux éléments clés des annotations, comme les phrases, les formes de surfaces (notées « *surface_form* ») ainsi que les annotations, tout en éliminant les balises XML non pertinentes pour notre étude.

Le processus de conversion a été réalisé en plusieurs étapes :

- Extraction des informations textuelles et des métadonnées depuis les balises XML
- Transformation de ces données en objets JSON structurés
- Validation pour s'assurer que toutes les informations essentielles telles que les identifiants de phrase, les formes de surfaces, et les annotations sémantiques, étaient correctement conservées

Cette conversion nous a permis de faciliter l'intégration des données dans l'API d'AMuSe-WSD, en rendant les fichiers plus légers, plus rapides à traiter, et plus compatibles avec les outils de traitement.

Nous avons appliqué AMuSe-WSD sur l'ensemble des phrases du corpus SemEval2013 pour obtenir des annotations en sens pour les noms présents.

Cette première étape, nous a permis de générer une couverture complète d'annotations sur les 306 phrases du corpus SemEval2013.

Le modèle AMuSe-WSD produit ces annotations en effectuant une tokenisation préalable, on obtient donc un total de 10 156 *tokens*, il y a 2 340 *tokens* annotés comme des *NOUN*, ce qui correspond aux noms communs (cf. la figure 3.2).

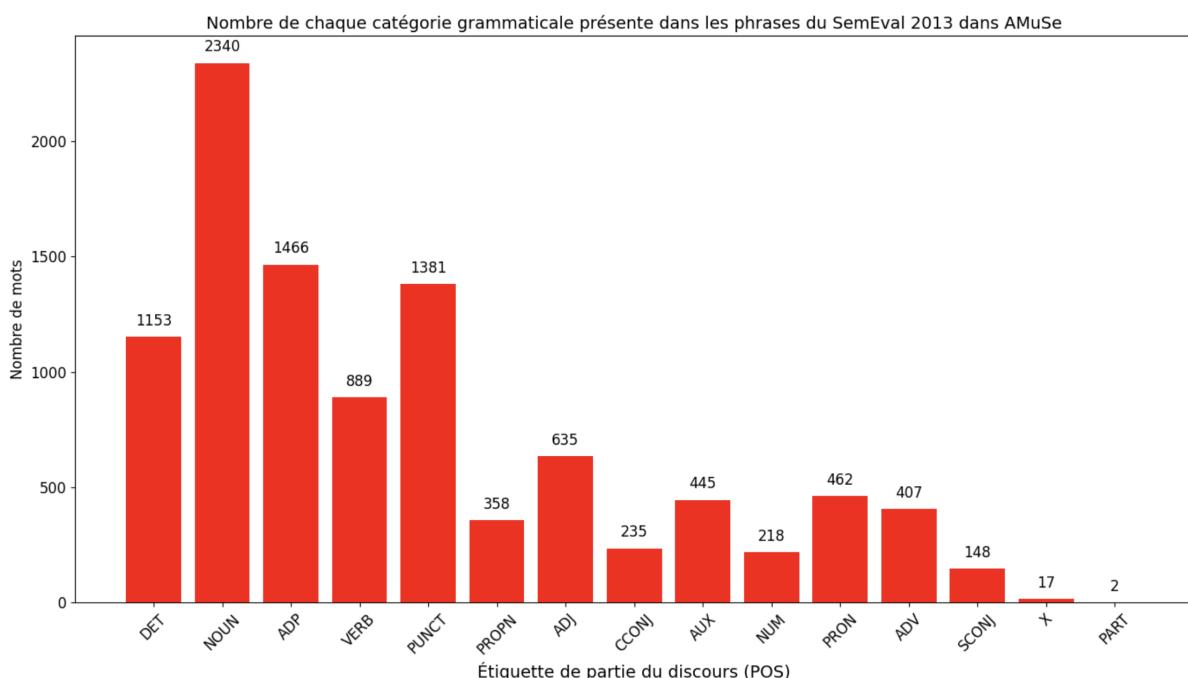


FIGURE 3.2 – Récupération des proportions de chaque catégorie grammaticale du corpus SemEval2013 à partir de l'API du modèle AMuSe-WSD

3.2 Annotation automatique du corpus SemEval2013 avec AMuSe-WSD

Pour évaluer le modèle AMuSe-WSD, nous avons utilisé son API avec le langage Python pour effectuer l'annotation automatique du corpus SemEval2013.

3.2.1 Comparaison de la tokenisation et de l'étiquetage en partie du discours

Dans un premier temps, nous avons examiné la tokenisation effectuée par AMuSe-WSD, en nous concentrant particulièrement sur les noms et les expressions polylexicales.

La tokenisation joue un rôle important dans la précision des annotations, car les expressions polylexicales (comme les noms composés ou les expressions figées) peuvent être traitées différemment selon les systèmes.

Par exemple, dans le corpus SemEval2013, les expressions polylexicales sont représentées avec des *underscores* (par exemple, nous retrouvons l'expression « secteur_pétrolier »), tandis qu'AMuSe-WSD les tokenise en deux unités distinctes, traitant « secteur » comme un nom et « pétrolier » comme un adjectif (cf. les figures 3.3).

Cette différence de traitement influence directement les annotations sémantiques obtenues.

Répartition des Expressions Polylexicales : Exclusives au fichier XML, Exclusives à AMuSe, et Communs aux deux sources

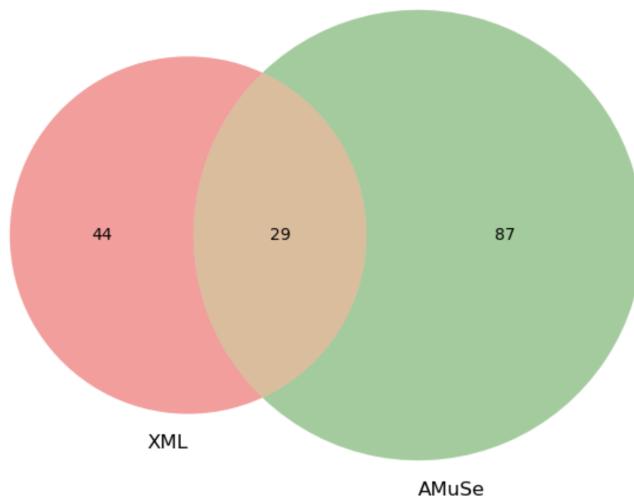


FIGURE 3.3 – Représentation des proportions des cas dans lesquels les expressions polylexicales (c.à.d. les noms écrits avec des *underscores*) apparaissent uniquement dans les données d'AMuSe-WSD, uniquement dans les données du SemEval2013 ou alors dans les deux jeux de données.

En analysant ces résultats, nous avons observé que certains noms présents uniquement dans les annotations générées par AMuSe-WSD correspondaient soit à des mots qui ne sont pas considérés comme des noms dans le corpus SemEval2013, soit à des expressions polylexicales dont la tokenisation diffère entre les deux systèmes (cf. figure 3.4).

Répartition des Noms : Exclusifs au fichier XML, Exclusifs à AMuSe, et Communs aux deux sources

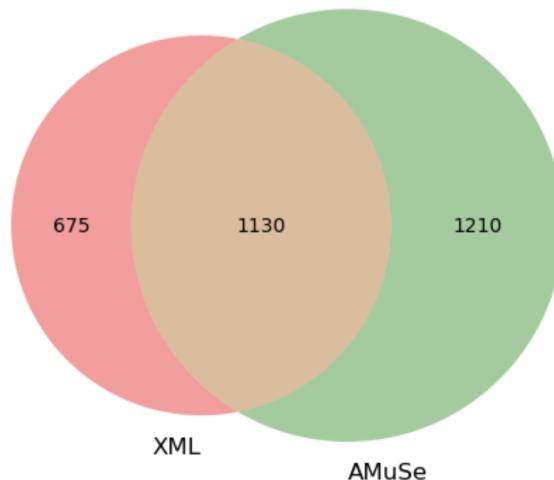


FIGURE 3.4 – Représentation des proportions des cas dans lesquels les noms apparaissent uniquement dans les données d’AMuSe-WSD, uniquement dans les données du SemEval2013 ou alors dans les deux jeux de données.

3.2.2 Comparaison des annotations en sens

Après avoir comparé la tokenisation, nous avons procédé à la comparaison des annotations sémantiques.

Pour ce faire, nous avons commencé par extraire tous les *sensekeys* associés à chaque lemme du corpus. Un *sensekey* est un identifiant textuel unique utilisé pour représenter un sens spécifique d’un mot dans WordNet. Chaque mot peut avoir plusieurs *sensekeys*, correspondant à ses différents sens. Nous avons également récupéré les noms associés à ces *sensekeys* afin de les lier aux termes annotés dans le corpus SemEval2013.

Ensuite, nous avons converti ces *sensekeys* en *synsets*, des identifiants numériques standards utilisés dans WordNet pour représenter des groupes de synonymes partageant un même sens.

Cette conversion était essentielle pour assurer une correspondance précise entre les annotations automatiques générées par AMuSe-WSD et celles des annotations *gold* de SemEval2013. En effet, les *synsets* fournissent une structure plus formelle et unifiée pour la représentation des sens, ce qui permet une comparaison plus rigoureuse et directe entre les différents jeux d’annotations.

Nous avons procédé à la comparaison des *synsets* générés par AMuSe-WSD avec ceux extraits des annotations *gold* du SemEval2013.

Cette comparaison a été réalisée de manière précise, en confrontant chaque *synset* lemme par lemme et phrase par phrase, afin de garantir une cohérence maximale dans l’évaluation des performances.

Ce processus minutieux a permis de détecter les correspondances exactes, ainsi que les divergences entre les annotations des deux systèmes, et ainsi de mesurer la précision du modèle dans la désambiguïsation lexicale (cf. les figures 4.3).

La comparaison détaillée des *synsets* a révélé qu’AMuSe-WSD a atteint une exactitude de 61,16 % avec 836 correspondances exactes sur un total de 1 367 annotations *gold* (sur

3.2 Annotation automatique du corpus SemEval2013 avec AMuSe-WSD

1 805 noms), et 531 divergences dues à des différences dans l'analyse des *synsets* (cf. 3.5).

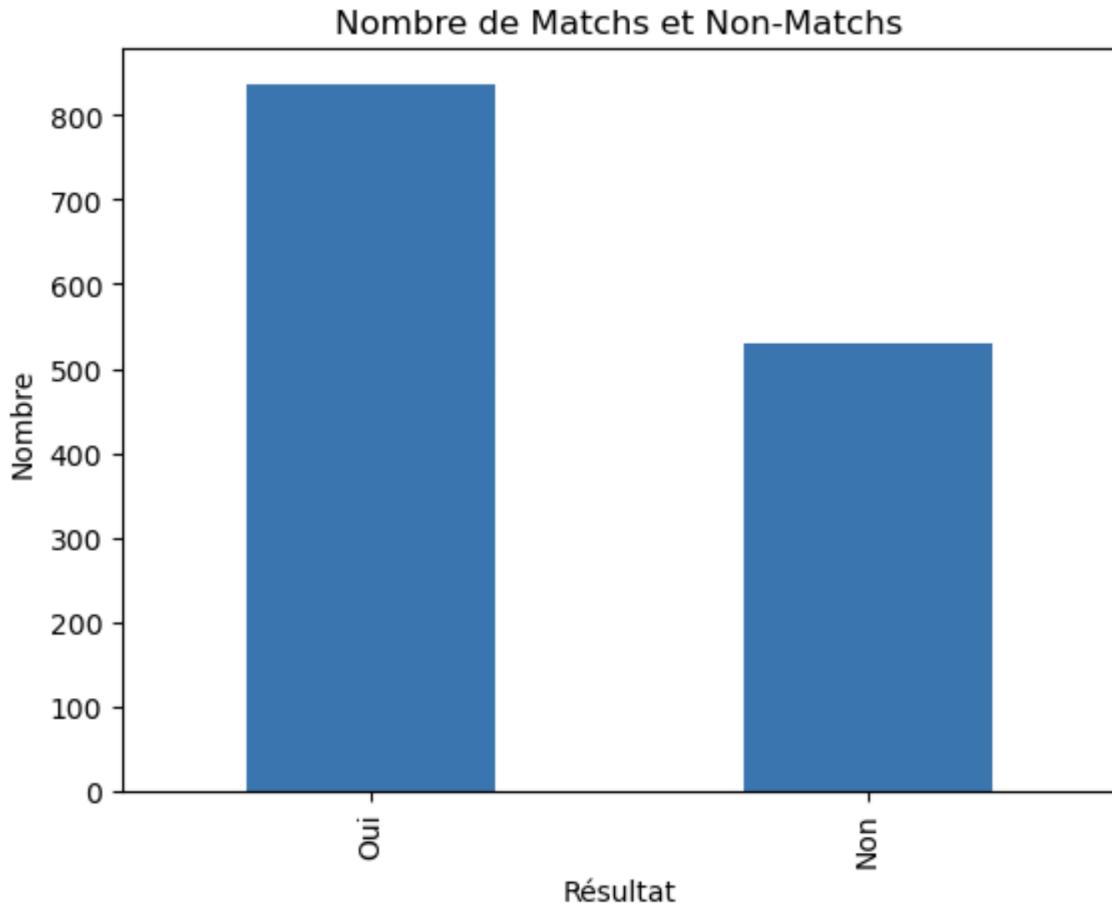


FIGURE 3.5 – Représentation graphique des *matches* de la comparaison des *synsets*

Ces différentes analyses, nous ont permis de dresser un portrait détaillé et exhaustif des performances d'AMuSe-WSD dans le contexte spécifique du corpus SemEval2013, fournissant ainsi des *insights*¹ pertinents pour proposer de nouvelles perspectives en vue d'optimiser la désambiguïsation lexicale.

1. Dans le contexte de ce mémoire, le terme « *insight* » désigne une observation ou une compréhension nouvelle qui émerge de l'analyse des données ou de la recherche.

Chapitre 4

Utilisation du modèle AMuSe-WSD sur le corpus FrSemCor

L'objectif de cette seconde partie de notre travail, est d'obtenir un corpus annoté en sens de manière semi-automatique sur le corpus du FrSemCor avec le modèle AMuSe-WSD. Nous avons conduit une procédure d'annotation manuelle en sens sur une partie du corpus, permettant ainsi d'évaluer la performance d'AMuSe-WSD.

4.1 Statistiques sur les annotations en *supersenses* des corpus SemEval2013 et FrSemCor

En premier lieu, nous avons choisi d'évaluer la répartition des *supersenses* présents dans le corpus SemEval2013 ainsi que dans le corpus FrSemCor.

Cette évaluation a pour objectif d'identifier les *supersenses* les plus fréquents et les plus pertinents pour orienter notre étude.

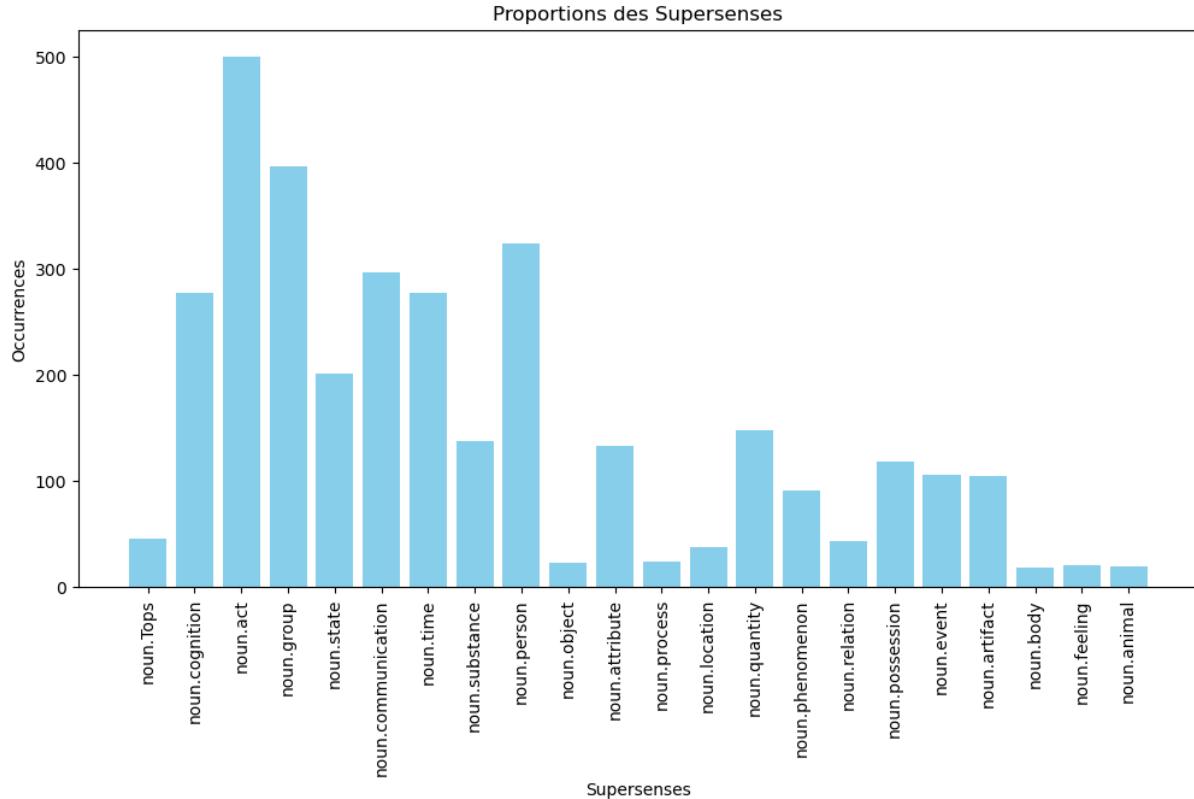
Pour évaluer la répartition des *Unique Beginners* dans le corpus SemEval2013, annoté en WSD, nous avons d'abord extrait les *Unique Beginners* directement à partir des annotations XML du corpus. Pour ce faire, nous avons converti le fichier XML en un format plus exploitable afin de faciliter l'extraction des données.
Ensuite, nous avons isolé les *Unique Beginners* associés à chaque annotation, puis compté leurs occurrences pour déterminer les catégories les plus récurrentes.

Cette analyse a permis d'identifier les *Unique Beginners* dominants dans SemEval2013. En parallèle, nous avons utilisé les données du corpus FrSemCor, fournies par Barque et al. (2020), pour comparer les proportions de *supersenses* et identifier les étiquettes communes aux deux corpus.

Corpus SemEval2013

Nous avons récupéré les *Unique Beginners* du corpus SemEval2013, ce qui nous a permis de comparer la répartition des *supersenses* entre les deux corpus.

La figure 4.1 montre les proportions des *Unique Beginners* trouvés dans le SemEval2013. En considérant que les classes les plus récurrentes sont celles avec au moins 200 occurrences, nous avons identifié 7 catégories principales : *noun.act*, *noun.person*, *noun.time*, *noun.cognition*, *noun.group*, *noun.communication*, et *noun.state*.


 FIGURE 4.1 – Proportions de *supersenses* trouvés dans le SemEval2013

Corpus FrSemCor

Le tableau 4.1 présente également les proportions des *supersenses* trouvés dans le corpus FrSemCor.

Selon les informations fournies par Barque et al. (2020), les *supersenses* les plus récurrents sont ceux qui apparaissent plus de 700 fois. Ces *supersenses* incluent : *Act*, *Person*, *Time*, *Cognition*, *Institution*, *Event*, et *Substance*. Le tableau 4.1 illustre la distribution des 20 *supersenses* les plus fréquents dans le corpus FrSemCor, ceux utilisés pour plus de 50 *tokens*.

En comparant les deux corpus, nous observons que les étiquettes dominantes communes sont :

- *Act*
- *Person*
- *Time*
- *Cognition*

Ces *supersenses* sont les plus fréquemment utilisés pour annoter les mots dans les deux corpus. Par conséquent, nous avons décidé de concentrer notre attention sur l'annotation de textes en utilisant ces quatre *supersenses*.

<i>Supersenses</i>	Nb lemma tokens	Nb lemma types
Act	2079	613
Person	1853	396
Time	1052	108
Cognition	906	252
Institution	735	155
Event	712	149
Substance	711	129
Quantity	687	93
State	621	160
Attribute	565	184
Artifact	530	236
Possession	366	71
Act/Cognition	314	97
GroupxPerson	222	54
Artifact/Cognition	213	51
Body	203	54
Object	159	56
Feeling	81	46
Part	67	28
Phenomenon	60	18

TABLEAU 4.1 – Distribution des 20 *supersenses* les plus fréquents dans le corpus annoté FrSemCor (= ceux utilisés pour plus de 50 tokens). (cf. Tableau 3 de Barque et al. (2020))

4.2 Annotations manuelles en sens sur le corpus FrSemCor

Comme mentionné précédemment, le corpus FrSemCor est annoté exclusivement en *supersenses* à l'aide de l'inventaire de sens WordNet (cf. la section 2.1.2). Pour évaluer l'efficacité du modèle AMuSe-WSD sur ce corpus, nous avons d'abord procédé à une annotation manuelle en *supersenses* du corpus FrSemCor.

Le corpus annoté est initialement au format CoNLL-U, un format tabulaire adapté aux annotations linguistiques, notamment pour les dépendances syntaxiques (cf. la figure 4.2 pour la présentation du format). Pour simplifier les annotations manuelles, nous avons choisi de travailler avec un format CSV, également tabulaire mais jugé plus adapté et plus simple d'utilisation pour nos besoins spécifiques.

Pour cette annotation, nous avons opté pour l'inventaire de sens BabelNet au lieu de WordNet, car BabelNet est mieux adapté à la langue française et fournit des *supersenses* plus précis pour les noms communs (cf. la section 2.2.2). Nous avons sélectionné le sens le plus approprié pour chaque nom en fonction du contexte, afin d'assurer une annotation pertinente et précise.

Initialement, nous avions prévu de nous concentrer sur les *supersenses* les plus fréquents dans le corpus, à savoir *Act*, *Time*, et *Person*. Cependant, nous avons observé que le *supersense* *Time* se réfère principalement à des dates et des chiffres, tandis que le *supersense* *Person* inclut un grand nombre d'entités nommées telles que des prénoms.

```

# sent_id = annodis.er_00001
# text = Gutenberg
1     Gutenberg      Gutenberg      NPP      s=p      Person

# sent_id = annodis.er_00002
# text = Cette exposition nous apprend que dès le XIIe siècle, à Dammarie-sur-Saulx, entre a
1     Cette     ce      DET      g=f|n=s|s=dem   *
2     exposition     exposition     NC      g=f|n=s|s=c      Act
3     nous     le      CLO      n=p|p=1|s=obj   *
4     apprend     apprendre     V      dl=apprendre|dm=ind|m=ind|n=s|p=3|t=pst   *
5     que     que      CS      s=s|void=y   *
6     dès     dès      P      _          *
7     le      le      DET      g=m|n=s|s=def   *
8     XIIe    XIIe     ADJ      s=ord   *
9     siècle    siècle     NC      def=y|g=m|n=s|s=c      Time
10    ,       ,       PONCT     s=w      *
11    à       à       P      _          *
12    Dammarie-sur-Saulx     Dammarie-sur-Saulx      NPP      g=m|n=s|s=p      Institution
13    ,       ,       PONCT     s=w      *
14    entre    entre     P      _          *
15    autres    autre     ADJ      n=p|s=ind   *
16    sites     site      NC      g=m|n=p|s=c      Artifact
17    ,       ,       PONCT     s=w      *
18    une     un      DET      g=f|n=s|s=ind   *
19    industrie    industrie     NC      g=f|n=s|s=c      Act
20    métallurgique    métallurgique     ADJ      n=s|s=qual   *
21    existait    exister     V      dl=exister|dm=ind|m=ind|n=s|p=3|t=impft   *
22    .       .       PONCT     s=s      *

```

FIGURE 4.2 – Extrait du corpus FrSemCor au format CoNLL-U ([Barque et al., 2020](#))

Ces catégories se sont révélées moins pertinentes pour l'annotation fine que nous souhaitions réaliser. Nous avons donc ajusté notre approche pour nous concentrer uniquement sur le *supersense Act*, qui s'est avéré le plus consistant et représentatif dans le corpus FrSemCor (cf. le tableau 4.1).

En effet, la majorité des noms annotés avec le *supersense Act* sont des noms communs, ce qui facilite leur annotation en sens (cf. la section 2.2.1).

Nous avons organisé notre tableau d'annotations comme suit, avec cinq colonnes : ***sent_id***, ***text***, ***word***, ***Supersense***, ***bn_id*** (identifiant BabelNet) et **commentaire** (cf. la figure 4.3).

- La colonne ***sent_id*** répertorie les identifiants de chaque phrase pour faciliter l'identification et la relecture des données.
- La colonne ***text*** contient les phrases analysées.
- La colonne ***word*** liste les noms à annoter.
- La colonne ***bn_id*** indique l'identifiant du sens dans BabelNet attribué au nom étudié.
- La colonne ***Supersense*** indique le ou les *supersenses* associés au nom.
- Enfin, la colonne **commentaire** permet de noter nos éventuelles hésitations ou remarques concernant nos annotations.

4.3 Observation des annotations manuelles effectuées sur le corpus FrSemCor

sent_id	text	word	Supersense	bn_id	Commentaire
annodis.er_00002	Cette exposition nous apprend que dès le XIIe siècle, à Damma exposition	Act		bn:00032225n	
annodis.er_00002	Cette exposition nous apprend que dès le XIIe siècle, à Damma industrie	Act		bn:16448017n	
annodis.er_00003	à peu près au même moment que Gutenberg inventait l'imprime imprimerie	Act		bn:00064453n	
annodis.er_00007	Amélioration de la sécurité	Amélioration	Act	bn:00046191n	
annodis.er_00008	Le maire a invité les membres du conseil à élaborer le programr amélioration	Act		bn:00046191n	
annodis.er_00010	La pose d'un panneau stop paraît être la formule la mieux adap pose	Act		bn:00046934n	Bn id : EN (pas traduction automa
annodis.er_00011	En délibérant, l'assemblée a accepté la proposition du maire et demande	Act		bn:00009665n	
annodis.er_00011	En délibérant, l'assemblée a accepté la proposition du maire et subvention	Act		bn:00074977n	
annodis.er_00011	En délibérant, l'assemblée a accepté la proposition du maire et répartition	Act		bn:00002911n	Bn id : EN (pas traduction automa
annodis.er_00013	La remise du lavoir abrite depuis quelques jours une exposition exposition	Act		bn:00032225n	
annodis.er_00013	La remise du lavoir abrite depuis quelques jours une exposition insertion	2;Act		bn:00046887n	bn:15228548n mais dans ce cor
annodis.er_00014	L'association a changé les décors et avec l'aide de plusieurs b aide	Act		bn:00006523n	
annodis.er_00014	L'association a changé les décors et avec l'aide de plusieurs b activité	Act		bn:00064608n	
annodis.er_00016	"Tout simplement", a précisé Roger Thiriot, "parce que l'histoire travail	Act		bn:00894737n	
annodis.er_00018	C'est donc toute la vie industrielle du bassin de Saint-Dizier, sa vie	Act		bn:00051049n	
annodis.er_00019	Un voyage étonnant où photos, réalisations, vieux outils, docum voyage	Act		bn:00078085n	
annodis.er_00019	Un voyage étonnant où photos, réalisations, vieux outils, docum réalisations	Act/Artifact		bn:00000724n	EN cependant je ne vois pas si or
annodis.er_00019	Un voyage étonnant où photos, réalisations, vieux outils, docum passage	Act		bn:00060883n	EN autre définition ne semble pas
annodis.er_00019	Un voyage étonnant où photos, réalisations, vieux outils, docum invasions	Act		bn:00047325n	

FIGURE 4.3 – Extrait de nos annotations manuelles du corpus FrSemCor

4.3 Observation des annotations manuelles effectuées sur le corpus FrSemCor

4.3.1 Calcul de l'accord inter-annotateurs

L'accord inter-annotateurs mesure le degré de concordance entre différents annotateurs lorsqu'ils identifient le sens d'un mot dans un contexte donné. Cette évaluation est cruciale pour garantir la précision et la cohérence des annotations entre les différents annotateurs.

Pour cette étude, nous avons calculé plusieurs coefficients d'accord inter-annotateurs afin d'évaluer la cohérence et la fiabilité des annotations réalisées dans notre corpus. Nous avons choisi d'utiliser le **coefficient de Kappa de Cohen**, noté κ , qui est une mesure statistique couramment utilisée pour quantifier le niveau d'accord entre deux annotateurs qui classent des éléments dans des catégories.

Le coefficient κ est particulièrement utile dans les situations où les décisions sont subjectives et où les catégories sont nominales. Ce score permet de mesurer la concordance observée au-delà de ce qui serait attendu par hasard.

L'accord observé, $Pr(a)$, est la proportion d'accord entre les annotateurs, calculée en divisant le nombre d'accords exacts par le nombre total d'annotations. Il représente la fréquence à laquelle les annotateurs se mettent d'accord sur la classification des éléments.

Le calcul du score de Kappa de Cohen a été automatisé grâce à l'utilisation de librairies spécialisées qui facilitent et standardisent le processus de mesure de l'accord inter-annotateurs.

L'outil a calculé la proportion d'accord observé sur un total de 100 annotations manuelles, avec une proportion d'accord $\text{Pr}(a)$ de

$$\text{Pr}(a) = \frac{\text{nombre d'accords exacts}}{\text{nombre total d'annotations}} = \frac{64}{100} = 0,64.$$

L'outil a également calculé la probabilité d'accord aléatoire, $\text{Pr}(e)$ en se basant sur les distributions de catégorie des annotations. En appliquant la formule de Kappa de Cohen en 4.4 :

$$\kappa = \frac{\text{Pr}(a) - \text{Pr}(e)}{1 - \text{Pr}(e)},$$

FIGURE 4.4 – Formule permettant de calculer le score Kappa de Cohen, où $\text{Pr}(a)$ est la proportion d'accord observé entre annotateurs (ou l'accord observé) et $\text{Pr}(e)$ la probabilité d'accord aléatoire.

L'interprétation des résultats obtenus avec le coefficient κ est importante pour comprendre le niveau d'accord. Le tableau 4.5 fournit une interprétation des scores κ en fonction de leur valeur :

κ	Interprétation
< 0	Désaccord
0,00 – 0,20	Accord très faible
0,21 – 0,40	Accord faible
0,41 – 0,60	Accord modéré
0,61 – 0,80	Accord fort
0,81 – 1,00	Accord presque parfait

FIGURE 4.5 – Tableau d'interprétation du Kappa de Cohen selon Wong et al. (2021). Les ordres de grandeur proposés ne font pas consensus dans la communauté scientifique, en raison de l'influence du nombre de catégories sur l'estimation : moins il y a de catégories, plus le κ est élevé.

Nous avons réalisé trois calculs d'accord inter-annotateurs avec les résultats suivants :

4.3 Observation des annotations manuelles effectuées sur le corpus FrSemCor

- Le premier accord inter-annotateur a donné un score de **0,71**, indiquant un **accord fort** ;
- Le second accord inter-annotateur a donné un score de **0,53**, indiquant un **accord modéré** ;
- Le troisième accord inter-annotateur a donné un score de **0,80**, indiquant un **accord fort**.

Pour améliorer la cohérence entre les annotateurs, nous avons d'abord annoté une quinzaine de phrases et avons ensuite analysé en détail les divergences observées entre cette première quinzaine d'annotations.

Nous avons organisé des sessions de discussion approfondies pour examiner en détail les divergences observées entre les résultats de l'accord inter-annotateurs. Cette analyse a impliqué une révision minutieuse des cas où les annotations avaient divergé, ainsi qu'une discussion sur les critères et les principes appliqués lors de l'annotation.

Nous avons identifié que l'utilisation de définitions en français, plutôt que des traductions de l'anglais, était une source majeure de divergence. Par conséquent, nous avons établi une règle principale stipulant que les définitions françaises, jugées plus pertinentes et précises, devaient être privilégiées lors de l'annotation. Cette règle visait à harmoniser les critères de sélection des sens et à réduire les ambiguïtés liées aux traductions.

En parallèle, nous avons également clarifié certains aspects des annotations, tels que les critères spécifiques pour déterminer le sens approprié dans des contextes particuliers, et avons mis en place un guide de référence commun pour les annotateurs. Ces mesures ont facilité une meilleure compréhension et une application plus uniforme des critères d'annotation.

Nous avons également noté que, entre chaque accord inter-annotateurs, de nouvelles entités ont été ajoutées aux échantillons. Ainsi, chaque accord a été réalisé sur des ensembles de données distincts, ce qui a permis d'observer les effets des ajustements méthodologiques sur des échantillons variés.

Cette concertation et ces ajustements méthodologiques ont eu un impact positif sur la qualité de l'accord inter-annotateur, comme le démontre l'augmentation significative du score d'accord inter-annotateur observé lors du dernier calcul. L'amélioration du score indique que les annotateurs ont atteint un consensus plus élevé et une meilleure cohérence dans leurs annotations, renforçant ainsi la fiabilité et la précision de notre processus d'annotation.

4.3.2 Comparaison des annotations générées par AMuSe-WSD avec les annotations *gold* sur le corpus FrSemCor

Après avoir terminé l'étape des annotations manuelles, il était indispensable d'évaluer la performance du modèle AMuSe-WSD en comparant ses annotations automatiques avec les nôtres. Pour réaliser cette comparaison, nous avons extrait les 56 phrases que nous avions annotées manuellement dans un fichier XML, afin de faciliter leur traitement par l'API du modèle.

Le fichier XML utilisé dans notre étude contient les annotations manuelles du corpus, structurées sous forme de paires clé-valeur. Chaque entrée comprend plusieurs champs essentiels : l’identifiant unique de la phrase (*sent_id*), le texte de la phrase complète (*text*), le mot annoté (*word*), le *supersense* attribué (*Supersense*), ainsi que l’identifiant BabelNet (*bn_id*) associé au sens du mot. Ce format permet de capturer à la fois le contexte textuel et les informations sémantiques des annotations.

L’analyse de l’output du modèle AMuSe-WSD a révélé que, pour chaque phrase, le modèle fournissait non seulement l’identifiant WordNet des noms annotés, mais aussi l’identifiant BabelNet. Cette double information s’est avérée particulièrement utile puisque nos annotations manuelles étaient basées sur BabelNet. Cela a permis une comparaison plus directe et précise. Nous avons donc choisi de concentrer notre évaluation sur les identifiants BabelNet plutôt que sur ceux de WordNet (cf. la figure 2.6).

```
[{"tokens": [{"index": 0, "text": "Le", "pos": "DET", "lemma": "le", "bnSynsetId": "0", "wnSynsetOffset": "0", "nltkSynset": "0"}, {"index": 1, "text": "groupe", "pos": "NOUN", "lemma": "groupe", "bnSynsetId": "bn:00041942n", "wnSynsetOffset": "31264n", "nltkSynset": "group.n.01"}, {"index": 2, "text": "des", "pos": "ADP", "lemma": "de", "bnSynsetId": "0", "wnSynsetOffset": "0", "nltkSynset": "0"}, {"index": 3, "text": "Nations_Unies", "pos": "NOUN", "lemma": "nations_unie", "bnSynsetId": "bn:00078931n", "wnSynsetOffset": "8295580n", "nltkSynset": "united_nat"}]
```

FIGURE 4.6 – Extrait de l’affichage des annotations AMuSe-WSD avec un exemple de l’identifiant BabelNet encadré en orange et de l’identifiant WordNet encadré en rose)

tableau_annotationeurs

sent_id	text	word	Supersense	Annotateur 1	Annotateur 2	
annodis.er_00002	Cette exposition nous apprend que dé	exposition	Act	bn:00032225n	bn:00032225n	✓
annodis.er_00002	Cette exposition nous apprend que dé	industrie	Act	bn:16448017n	bn:16448017n	✓
annodis.er_00003	à peu près au même moment que Gut	imprimerie	Act	bn:13604369n	bn:00064453n	✗
annodis.er_00007	Amélioration de la sécurité	Amélioration	Act	bn:00046191n	bn:00046191n	✓
annodis.er_00008	Le maire a invité les membres du cons	amélioration	Act	bn:00046191n	bn:00046191n	✓
annodis.er_00010	La pose d'un panneau stop paraît être	pose	Act	bn:00046934n	bn:00046934n	✓
annodis.er_00011	En délibérant, l'assemblée a accepté	demande	Act	bn:00009665n	bn:00009665n	✓
annodis.er_00011	En délibérant, l'assemblée a accepté	subvention	Act	bn:00074977n	bn:00074977n	✓
annodis.er_00011	En délibérant, l'assemblée a accepté	répartition	Act	bn:00002911n	bn:00002911n	✓
annodis.er_00013	La remise du lavoir abrite depuis quelk	exposition	Act	bn:00032225n	bn:00032225n	✓
annodis.er_00013	La remise du lavoir abrite depuis quelk	insertion	2:Act	bn:00046887n	bn:00046887n	✓
annodis.er_00014	L'association a changé les décors et a	aide	Act	bn:16784964n	bn:00006523n	✗
annodis.er_00014	L'association a changé les décors et a	activité	Act	bn:15840810n	bn:00064608n	✗
annodis.er_00016	"Tout simplement", a précisé Roger T	travail	Act	bn:00894737n	bn:00894737n	✓
annodis.er_00018	C'est donc toute la vie industrielle du	vie	Act	bn:00051048n	bn:00051049n	✗
annodis.er_00019	Un voyage étonnant où photos, réalis	voyage	Act	bn:00078085n	bn:00078085n	✓
annodis.er_00019	Un voyage étonnant où photos, réalis	réalisations	Act/Artifact	bn:00000724n	bn:00000724n	✓
annodis.er_00019	Un voyage étonnant où photos, réalis	passage	Act	bn:00060883n	bn:00060883n	✓
annodis.er_00019	Un voyage étonnant où photos, réalis	invasions	Act	bn:00047325n	bn:00047325n	✓
annodis.er_00019	Un voyage étonnant où photos, réalis	labeur	Act	bn:00049571n	bn:00028232n	✗
annodis.er_00019	Un voyage étonnant où photos, réalis	travail	Act	bn:00894737n	bn:00894737n	✓
annodis.er_00020	Cette exposition, comme devait concl	exposition	Act	bn:00032225n	bn:00032225n	✗
annodis.er_00020	Cette exposition, comme devait concl	tissu	2:GroupxAct	UNK	UNK	✓

FIGURE 4.7 – Extrait de notre tableau d’annotation avec les différents annotateurs et les correspondances ou non des *synsets*

Pour garantir la cohérence des comparaisons entre les identifiants BabelNet fournis

4.3 Observation des annotations manuelles effectuées sur le corpus FrSemCor

par AMuSe-WSD et nos propres annotations, nous avons procédé à une concertation entre annotateurs. Nous avons discuté des divergences et décidé, pour chaque mot, du sens que nous jugions le plus approprié dans le contexte, afin d'harmoniser les annotations sur une centaine de cas. Une fois cet accord atteint, nous avons déployé AMuSe-WSD sur les 56 phrases annotées et avons procédé à l'extraction des identifiants BabelNet générés par le modèle.

La comparaison des identifiants BabelNet entre nos annotations et celles produites par AMuSe-WSD a été réalisée à l'aide de scripts Python, et nous avons obtenu un taux d'exactitude de 40,86 %. Pour illustrer ces résultats, nous avons créé un tableau récapitulatif des annotations de chaque annotateur (cf. la figure 4.8) ainsi qu'une représentation graphique des correspondances entre les *synsets* BabelNet (cf. la figure 4.9).

resultats_comparaison

ID Phrase	Mot	Synset Annotation Manuelle	Synset AMuSe	Match Status
annodis.er_00002	exposition	bn:00032225n	bn:00032225n	Match
annodis.er_00002	industrie	bn:16448017n	bn:00046576n	No Match
annodis.er_00003	imprimerie	bn:13604369n	bn:00064454n	No Match
annodis.er_00007	amélioration	bn:00046191n	bn:00046191n	Match
annodis.er_00008	amélioration	bn:00046191n	bn:00046191n	Match
annodis.er_00010	pose	bn:00046934n	bn:00046934n	Match
annodis.er_00011	demande	bn:00009665n	bn:00061826n	No Match
annodis.er_00011	subvention	bn:00074977n	bn:00074977n	Match
annodis.er_00011	répartition	bn:00002911n	bn:00002911n	Match
annodis.er_00013	exposition	bn:00032225n	bn:00032225n	Match
annodis.er_00013	insertion	bn:00046887n	bn:00046887n	Match
annodis.er_00014	aide	bn:16784964n	bn:00002155n	No Match
annodis.er_00014	activité	bn:15840810n	bn:00001128n	No Match
annodis.er_00016	travail	bn:00894737n	bn:00018756n	No Match
annodis.er_00018	vie	bn:00051048n	bn:00051045n	No Match
annodis.er_00019	voyage	bn:00078085n	bn:00078085n	Match
annodis.er_00019	réalisations	bn:00000724n	bn:00000724n	Match
annodis.er_00019	passage	bn:00060883n	bn:00017678n	No Match
annodis.er_00019	invasions	bn:00047325n	bn:00047325n	Match
annodis.er_00019	labeur	bn:00049571n	bn:00028232n	No Match
annodis.er_00019	travail	bn:00894737n	bn:00018756n	No Match

FIGURE 4.8 – Résultats des comparaisons faites entre les *synsets* BabelNet de nos annotations et ceux générés par AMuSe-WSD

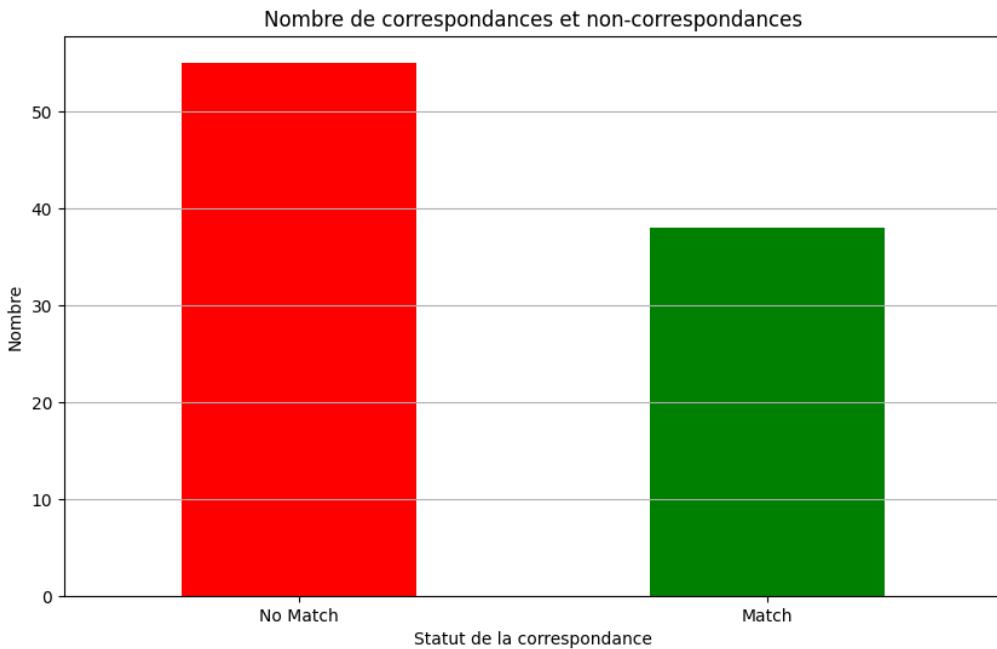


FIGURE 4.9 – Représentation graphique des *matches* de la comparaison des *synsets* BabelNet

L'exactitude de 40,86 % peut sembler relativement faible, mais en examinant le tableau des comparaisons, nous constatons que plusieurs noms n'ont pas été correctement annotés par AMuSe-WSD, pour diverses raisons :

1. Une tokenisation différentes des mots : Par exemple, dans le corpus FrSemCor, le préfixe « auto » et le nom « financement » sont tokenisés comme deux unités distinctes, tandis qu'AMuSe-WSD traite « auto-financement » comme une seule unité. Cela empêche donc une comparaison directe.
2. Différences dans les parties du discours : Certains mots, comme « rendez-vous », sont annotés comme des verbes par AMuSe-WSD alors qu'ils sont des noms dans le corpus FrSemCor. De même, « Billard » est annoté comme un nom propre dans AMuSe-WSD, créant ainsi des divergences.

En outre, pour certains mots, le contexte n'est pas correctement interprété par AMuSe-WSD, ce qui conduit à des différences dans les annotations. Il est probable qu'avec un échantillon de noms plus important, l'exactitude aurait été plus élevée.

Conclusion et Perspectives

Cette recherche a permis d'approfondir notre compréhension de la complexité de la désambiguïsation lexicale en français, tout en évaluant les performances du modèle **AMuSe-WSD** dans des tâches d'annotation sémantique. L'évaluation du modèle sur le corpus **SemEval2013** a révélé une exactitude de 61 %, un résultat qui témoigne de la capacité du modèle à accomplir des tâches complexes de désambiguïsation lexicale malgré la diversité et la difficulté du corpus. Cette performance est notable compte tenu du nombre élevé de labels possibles pour chaque mot, mais elle indique également un potentiel d'amélioration, notamment dans la gestion des ambiguïtés sémantiques multiples et des expressions polylexicales.

En parallèle, nous avions pour objectif d'améliorer le modèle **AMuSe-WSD** avant de l'appliquer à un autre corpus, **FrSemCor**, pour effectuer une annotation sémantique plus spécifique au français. Cependant, nous n'avons pas pu atteindre cet objectif en raison de contraintes de temps et de la complexité des étapes d'annotation manuelle. Nous avons donc concentré nos efforts sur l'annotation de **FrSemCor**, une tâche particulièrement fastidieuse nécessitant l'attribution d'un ID **BabelNet** à chaque mot annoté, tout en harmonisant les annotations entre plusieurs annotateurs.

Le processus d'annotation s'est déroulé en plusieurs phases et a intégré des accords inter-annotateurs successifs pour améliorer la cohérence des annotations. Trois échantillons distincts ont été soumis à cet accord inter-annotateurs, et les résultats ont montré une amélioration progressive à chaque phase, ce qui reflète une meilleure compréhension partagée des annotations à mesure que le travail avançait. Ce travail méticuleux d'harmo-nisation a permis d'enrichir le corpus **FrSemCor** avec des annotations *gold*, à la fois pour la désambiguïsation lexicale et pour les *supersenses*, ce qui est crucial pour les futures étapes de réévaluation des modèles.

L'annotation manuelle, bien qu'exigeante, a révélé des aspects importants de la désam-biguïsation lexicale. Elle a notamment mis en lumière les défis que posent les mots ambigus dans des contextes complexes, et comment l'annotation humaine peut encore surpasser les modèles automatiques dans certains cas. Ce travail nous a permis de mieux cerner les obstacles à l'automatisation de cette tâche, tout en identifiant des pistes d'amélioration pour les futures évolutions du modèle **AMuSe-WSD**.

Il est également important de noter que la performance du modèle sur le corpus **FrSemCor** est restée en deçà des attentes, principalement en raison du manque d'opti-misation spécifique au contexte du français et des contraintes temporelles. Cela indique que des améliorations sont nécessaires, et nous avons identifié plusieurs axes à explorer pour renforcer les capacités du modèle.

Parmi ces pistes d'amélioration, la gestion des expressions polylexicales ressort comme un élément clé. En effet, le traitement des locutions figées, des expressions idiomatiques et des termes techniques, souvent mal tokenisés ou mal interprétés par le modèle, pourrait être grandement amélioré par des algorithmes de tokenisation plus robustes et adaptés au français. De plus, l'augmentation et la diversification du jeu de données d'entraînement apparaissent comme des stratégies nécessaires pour offrir au modèle une meilleure couverture lexicale et contextuelle, augmentant ainsi sa capacité à traiter un éventail plus large de contextes linguistiques.

Un autre aspect clé concerne les accords inter-annotateurs que nous avons réalisés tout au long du processus d'annotation. Les résultats de ces accords ont montré une amélioration progressive, passant d'un taux initial relativement bas à un consensus de plus en plus fort entre les annotateurs. Cela démontre que l'amélioration des méthodes d'annotation et la formation des annotateurs jouent un rôle crucial dans la précision des corpus annotés, ce qui aura un impact direct sur les modèles d'apprentissage supervisé qui seront utilisés dans le futur.

En conclusion, cette étude a mis en évidence plusieurs aspects essentiels de la désambiguïsation lexicale et des défis spécifiques à l'annotation sémantique en français. Bien que le modèle **AMuSe-WSD** ait montré des performances satisfaisantes sur le corpus **SemEval2013**, il reste encore des améliorations à apporter, notamment pour la gestion des expressions complexes et pour l'optimisation du modèle dans des contextes linguistiques plus spécialisés, tels que ceux rencontrés dans **FrSemCor**. Le travail d'annotation manuel a non seulement enrichi le corpus, mais il a aussi fourni une compréhension plus fine des enjeux de la désambiguïsation, ouvrant ainsi la voie à de futures améliorations techniques.

Ces perspectives d'amélioration sont détaillées dans la section suivante, où nous explorerons comment augmenter la robustesse du modèle, affiner la gestion des données complexes, et optimiser l'algorithme de désambiguïsation pour des résultats plus précis et adaptés à une diversité de contextes linguistiques.

Voici certaines perspectives d'amélioration détaillées que nous pensons envisageables pour encore affiner cette étude.

Intégration d'une détection des *supersenses* au code source du modèle AMuSe-WSD

Ajouter un mécanisme de détection des *supersenses* au modèle **AMuSe-WSD** pourrait simplifier et renforcer le processus de désambiguïsation en utilisant les *supersenses* pour offrir une vue d'ensemble plus claire des significations possibles. En effet, en incorporant des catégories sémantiques générales en complément des annotations de sens, le modèle serait mieux équipé pour gérer les variations contextuelles et améliorer la précision des annotations.

Pour mettre en œuvre cette fonctionnalité, il serait d'abord nécessaire d'enrichir les données d'entraînement avec des annotations manuelles en *supersenses*, ou de complé-

ter les annotations manuelles en sens existantes. Dans notre cas, il aurait été bénéfique d'ajouter des annotations en sens lexicaux au corpus `FrSemCor`, qui est déjà annoté en *supersenses*.

Ensuite, le code source de `AMuSe-WSD` devrait être modifié pour intégrer un module capable de détecter ces *supersenses*. Cette modification devrait être testée afin d'évaluer son impact sur les performances globales du modèle, en comparant les résultats obtenus avec ceux précédemment enregistrés.

L'intégration de cette détection directement dans le code source du modèle `AMuSe-WSD` pourrait considérablement améliorer sa capacité à gérer la complexité sémantique, fourniissant ainsi des annotations plus précises et mieux adaptées aux contextes variés.

Amélioration de la tokenisation par `AMuSe-WSD`

L'étude a mis en lumière une difficulté pour `AMuSe-WSD` à gérer les expressions poly-lexicales, notamment les locutions figées et les termes techniques. Ces expressions, qui ont un sens propre différent de la somme des mots individuels, sont souvent mal analysées par le modèle.

Pour remédier à ce problème, il serait pertinent d'intégrer des dictionnaires d'expressions poly-lexicales ou de former le modèle sur des corpus annotés spécifiquement avec ces expressions, comme cela a été fait pour les corpus `SemEval2013` et `FrSemCor`.

Augmentation des données d'entraînement du modèle

Enrichir le modèle avec des données d'entraînement plus vastes et diversifiées permettrait d'améliorer ses performances. Il serait bénéfique d'augmenter la couverture lexicale en ajoutant des termes et expressions issus de différents domaines spécialisés (médecine, droit, informatique, etc.).

L'impact de cette augmentation serait évalué à l'aide de métriques telles que la précision, le rappel, l'exactitude et la F1-mesure, en comparant les résultats avec ceux obtenus avant l'augmentation des données.

Annexes

Tableau d'annotations manuelles sur le corpus FrSemCor

sent_id	text	word	Supersense	bn_id
annodis.er_00002	Cette exposition nous apprend que dès le XIIe siècle, à Dammarie-sur-Saulx, entre autres sites, un	exposition	Act	bn:00032225n
annodis.er_00002	Cette exposition nous apprend que dès le XIIe siècle, à Dammarie-sur-Saulx, entre autres sites, un	industrie	Act	bn:16448017n
annodis.er_00003	à peu près au même moment que Gutenberg inventait l'imprimerie, Gillet Bonnemire créait en 1450	imprimerie	Act	bn:13604369n
annodis.er_00007	Amélioration de la sécurité	Amélioration	Act	bn:00046191n
annodis.er_00008	Le maire a invité les membres du conseil à élaborer le programme d'amélioration de la voirie comm	amélioration	Act	bn:00046191n
annodis.er_00010	La pose d'un panneau stop paraît être la formule la mieux adaptée pour assurer la sécurité des usa	pose	Act	bn:00046934n
annodis.er_00011	En délibérant, l'assemblée a accepté la proposition du maire et l'a chargé de faire établir par les se	demande	Act	bn:00009665n
annodis.er_00011	En délibérant, l'assemblée a accepté la proposition du maire et l'a chargé de faire établir par les se	subvention	Act	bn:00074977n
annodis.er_00011	En délibérant, l'assemblée a accepté la proposition du maire et l'a chargé de faire établir par les se	répartition	Act	bn:00002911n
annodis.er_00013	La remise du lavoir abrite depuis quelques jours une exposition qui a été inaugurée par Roger Thiri	exposition	Act	bn:00032225n
annodis.er_00013	La remise du lavoir abrite depuis quelques jours une exposition qui a été inaugurée par Roger Thiri	insertion	2;Act	bn:00046887n
annodis.er_00014	L'association a changé les décors et avec l'aide de plusieurs bénévoles, établi différents tableaux s aide	activité	Act	bn:16784964n
annodis.er_00014	L'association a changé les décors et avec l'aide de plusieurs bénévoles, établi différents tableaux s aide	act	bn:00064608n	
annodis.er_00016	"Tout simplement", a précisé Roger Thiriot, "parce que l'histoire du travail industriel est, ici, une lon	travail	Act	bn:00894737n
annodis.er_00018	C'est donc toute la vie industrielle du bassin de Saint-Dizier, sans oublier les papeteries de Jeand'	vie	Act	bn:00051048n
annodis.er_00019	Un voyage étonnant où photos, réalisations, vieux outils, documents anciens, permettront de mesu	voyage	Act	bn:00078085n
annodis.er_00019	Un voyage étonnant où photos, réalisations, vieux outils, documents anciens, permettront de mesu	réalisations	Act/Artifact	bn:00000724n
annodis.er_00019	Un voyage étonnant où photos, réalisations, vieux outils, documents anciens, permettront de mesu	passage	Act	bn:00060883n
annodis.er_00019	Un voyage étonnant où photos, réalisations, vieux outils, documents anciens, permettront de mesu	invasions	Act	bn:00047325n
annodis.er_00019	Un voyage étonnant où photos, réalisations, vieux outils, documents anciens, permettront de mesu	labeur	Act	bn:00049571n
annodis.er_00019	Un voyage étonnant où photos, réalisations, vieux outils, documents anciens, permettront de mesu	travail	Act	bn:00894737n
annodis.er_00020	Cette exposition, comme devait conclure Roger Thiriot, "n'a d'autre ambition que d'apporter un mo	exposition	Act	bn:00032225n
annodis.er_00020	Cette exposition, comme devait conclure Roger Thiriot, "n'a d'autre ambition que d'apporter un mo	tissu	2:GroupxAct	UNK
annodis.er_00022	Ouverture tous les jours sauf le lundi de 14h30 à 18h.	Ouverture	Act	bn:00059095n
annodis.er_00024	Au cours de la cérémonie d'inauguration.	cérémonie	Act	bn:00017345n
annodis.er_00024	Au cours de la cérémonie d'inauguration.	inauguration	Act	bn:00046247n
annodis.er_00026	Après avoir coupé le ruban qui en marque symboliquement l'entrée, le maire et la directrice ont co	visite	Act	bn:00080111n
annodis.er_00027	Dans son intervention, M. Soyer a fait l'historique de l'école maternelle, qui existe à Vignot depuis	intervention	Act+Cognition	bn:00047061n
annodis.er_00027	Dans son intervention, M. Soyer a fait l'historique de l'école maternelle, qui existe à Vignot depuis	historique	Act/Cognition	bn:00044268n
annodis.er_00030	Une réflexion commune est menée avec les enseignants et les délégués de parents d'élèves, sous	réflexion	Act/Cognition	bn:00017339n
annodis.er_00030	Une réflexion commune est menée avec les enseignants et les délégués de parents d'élèves, sous	conduite	Act	bn:00009654n
annodis.er_00031	Après une année d'études, la conduite des travaux est menée par le cabinet Cadel et son associé,	études	Act	bn:00074790n
annodis.er_00031	Après une année d'études, la conduite des travaux est menée par le cabinet Cadel et son associé,	conduite	Act	bn:00009654n
annodis.er_00031	Après une année d'études, la conduite des travaux est menée par le cabinet Cadel et son associé,	travaux	Act	bn:00046568n
annodis.er_00031	Après une année d'études, la conduite des travaux est menée par le cabinet Cadel et son associé,	contrôle	Act	bn:06572017n
annodis.er_00032	Le coût des bâtiments s'élève à 2.700.000 F, dont 40% de subvention de l'Etat ; 20% d'auto-financ	subvention	Act	bn:00074977n
annodis.er_00032	Le coût des bâtiments s'élève à 2.700.000 F, dont 40% de subvention de l'Etat ; 20% d'auto-financ	financement	Act/Possession	bn:00034545n
annodis.er_00032	Le coût des bâtiments s'élève à 2.700.000 F, dont 40% de subvention de l'Etat ; 20% d'auto-financ	emprunt	Act/Possession	bn:00051693n
annodis.er_00034	M. Soyer a adressé ses remerciements aux collectivités participantes, et en particulier à l'inspecteu	remerciements	Act/Cognition	bn:00041465n
annodis.er_00034	M. Soyer a adressé ses remerciements aux collectivités participantes, et en particulier à l'inspecteu	création	Act	bn:00023652n
annodis.er_00034	M. Soyer a adressé ses remerciements aux collectivités participantes, et en particulier à l'inspecteu	nomination	Act	bn:00005113n
annodis.er_00034	M. Soyer a adressé ses remerciements aux collectivités participantes, et en particulier à l'inspecteu	mise	2:Act	UNK
annodis.er_00034	M. Soyer a adressé ses remerciements aux collectivités participantes, et en particulier à l'inspecteu	rentrée	Act	bn:15615843n
annodis.er_00035	M. Dumez, président du conseil général, a parlé d'une "réalisation apte à l'épanouissement des pe	réalisation	Act	bn:00000724n
annodis.er_00035	M. Dumez, président du conseil général, a parlé d'une "réalisation apte à l'épanouissement des pe	éducation	Act	bn:00026980n
annodis.er_00036	L'inspecteur d'académie a remercié la municipalité d'avoir mené à bien son projet en ayant associe	projet	Act/Cognition	bn:00064665n
annodis.er_00036	L'inspecteur d'académie a remercié la municipalité d'avoir mené à bien son projet en ayant associe	conception	Act	bn:25664677n
annodis.er_00037	Quant au sous-préfet, il apprécie l'énergie dépensée pour une telle réalisation.	réalisation	Act	bn:00000724n
annodis.er_00038	Il pense que l'espace, les couleurs, le silence d'un tel lieu ne peuvent qu'être un élément d'harmon	élément	PartxAct	bn:00027912n
annodis.er_00038	Il pense que l'espace, les couleurs, le silence d'un tel lieu ne peuvent qu'être un élément d'harmon	vie	Act	bn:00051048n
annodis.er_00038	Il pense que l'espace, les couleurs, le silence d'un tel lieu ne peuvent qu'être un élément d'harmon	projets	Act/Cognition	bn:00064665n
annodis.er_00040	Ce sont finalement les élèves de la classe de CM2 de Jacky Hedin qui sont venus mettre la main a	jeu	Act	bn:00037180n
annodis.er_00041	Une seconde opération se déroulait en parallèle sur le territoire de la commune, avec un groupe d'u	opération	Act	bn:00054973n
annodis.er_00042	Une opération destinée aussi à faciliter la remontée et la fraye des truites vers la rivière.	opération	Act	bn:00054973n
annodis.er_00042	Une opération destinée aussi à faciliter la remontée et la fraye des truites vers la rivière.	remontée	Act	bn:00175840n
annodis.er_00042	Une opération destinée aussi à faciliter la remontée et la fraye des truites vers la rivière.	fraye	Act	UNK
annodis.er_00043	La journée de nettoyage de l'environnement placée sous la bannière du "Printemps de l'environnem	journée	Act	bn:00025419n
annodis.er_00043	La journée de nettoyage de l'environnement placée sous la bannière du "Printemps de l'environnem	nettoyage	Act	bn:00019647n
annodis.er_00043	La journée de nettoyage de l'environnement placée sous la bannière du "Printemps de l'environnem	Printemps	1:Act;Time	bn:00073636n
annodis.er_00048	Ceux d'Ancerville ont participé à l'événement en faisant découvrir leur savoir-faire et leurs recettes	événement	Act	bn:14035017
annodis.er_00049	Les élèves de la classe de CE1 de l'école Notre-Dame ont pris part aux festivités.	festivités	Act	bn:00060836n

annodis.er_00050	Ils ont été reçus à la boulangerie Leroy pour visiter le fournil et surtout pétrir la pâte afin de confect	goûter	Act	bn:00374936n
annodis.er_00055	Football	Football	Act	bn:00006547n
annodis.er_00056	Une seule rencontre est au programme des licenciés du FC aujourd'hui.	rencontre	Act	bn:00041948n
annodis.er_00058	Une journée du championnat de promotion de première division départementale.	championnat	Act	bn:00017615n
annodis.er_00058	Une journée du championnat de promotion de première division départementale.	promotion	Act	bn:03493047n
annodis.er_00059	Cyclisme	Cyclisme	Act	bn:00024734n
annodis.er_00060	Les cyclistes et vététistes peuvent se réunir ce matin, à 9h, place Jacques-Bailleurs, à l'occasion d'une sortie	sortie	Act	bn:00032243n
annodis.er_00060	Les cyclistes et vététistes peuvent se réunir ce matin, à 9h, place Jacques-Bailleurs, à l'occasion d'un entraînement	entraînement	Act	bn:00028736n
annodis.er_00061	Cet entraînement sera renouvelé demain, aux mêmes horaires.	entraînement	Act	bn:00028736n
annodis.er_00062	Billard	Billard	Act	bn:00071206n
annodis.er_00064	Tir	Tir	Act	bn:02217337n
annodis.er_00065	Les tireurs de la Vaux-Racine ont rendez-vous à partir de 9h, sur les pas de tir, à l'occasion d'une séance de tir	séance	Act	bn:02217337n
annodis.er_00065	Les tireurs de la Vaux-Racine ont rendez-vous à partir de 9h, sur les pas de tir, à l'occasion d'une séance d'entraînement	entraînement	Act	bn:00070690n
annodis.er_00065	Les tireurs de la Vaux-Racine ont rendez-vous à partir de 9h, sur les pas de tir, à l'occasion d'une séance d'entraînement	entraînement	Act	bn:00028736n
annodis.er_00066	Aviron	Aviron	Act	bn:00068428n
annodis.er_00067	Une sortie nautique d'entraînement est organisée aujourd'hui, à partir de 9h, pour les rameurs du club.	sortie	Act	bn:00032243n
annodis.er_00067	Une sortie nautique d'entraînement est organisée aujourd'hui, à partir de 9h, pour les rameurs du club.	entraînement	Act	bn:00028736n
annodis.er_00068	Rendez-vous au local du club.	Rendez-vous	Act	bn:00067110n
annodis.er_00073	Il avait épousé Denise Pierrejean le 26 octobre 1974 et de leur union, sont nés une fille et deux garçons.	union	Act/State	bn:00053518n
annodis.er_00075	Guy Hosneld était une personne très connue dans la commune et aux alentours, pour son grand dévouement.	participation	Act	bn:27268958n
annodis.er_00075	Guy Hosneld était une personne très connue dans la commune et aux alentours, pour son grand dévouement.	vie	Act	bn:00051048n
annodis.er_00076	Il a assuré la présidence de la société de pêche "La Gaule vidusienne" durant une dizaine d'années.	présidence	Act	bn:00064232n
annodis.er_00076	Il a assuré la présidence de la société de pêche "La Gaule vidusienne" durant une dizaine d'années.	pêche	Act	bn:00034862n
annodis.er_00077	à l'âge de 12 ans, il était entré au LAS handball où il pratiqua durant près de vingt-cinq ans.	handball	Act	bn:25312443n
annodis.er_00078	Sapeur-pompier volontaire, il avait pris la responsabilité de la formation des premiers cadets sapeurs-pompiers.	formation	Act	bn:00026980n
annodis.er_00079	Guy affectionnait beaucoup le jardinage, loisir partagé avec Denise, son épouse, qu'il a eu la doule	jardinage	Act	bn:00037332n
annodis.er_00079	Guy affectionnait beaucoup le jardinage, loisir partagé avec Denise, son épouse, qu'il a eu la douleur	loisir	Act	bn:00027865n
annodis.er_00080	Il laisse à sa famille et amis, le souvenir d'un homme dévoué, généreux, jovial, toujours prêt à rendre service.	service	Act	bn:00070652n
annodis.er_00081	Les obsèques seront célébrées demain, à 15h, en l'église Notre-Dame de Void-Vacon.	obsèques	Act	bn:00036855n
annodis.er_00085	La première exposition avicole de Belfort date de 1922.	exposition	Act	bn:00032225n
annodis.er_00088	Pour être précis, c'est autour de 1950 que la société d'aviculture de Belfort a vraiment pris sa réelle	aviculture	Act	bn:00216945n
annodis.er_00088	Pour être précis, c'est autour de 1950 que la société d'aviculture de Belfort a vraiment pris sa réelle	présidence	Act	bn:00064232n
annodis.er_00090	C'est le cas de ce brave Joseph Bari, toujours présent au comité, vedette de la colombophilie jusqu'à	comité	Act	bn:00021015n
annodis.er_00090	C'est le cas de ce brave Joseph Bari, toujours présent au comité, vedette de la colombophilie jusqu'à	colombophilie	Act	bn:01650535n
annodis.er_00091	à l'époque, ce cher "Jo" élevait jusqu'à 120 pigeons voyageurs, et les spécialistes n'ont pas oublié son retour.	retour	Act	bn:00067495n
annodis.er_00093	Heureusement, comme les autres amis de Claude Simon, il adore consacrer son temps à cette très	exposition	Act	bn:00032225n
annodis.er_00093	Heureusement, comme les autres amis de Claude Simon, il adore consacrer son temps à cette très	chants	Act	bn:27267991n
annodis.er_00097	Une manoeuvre menée avec maestria.	manoeuvre	Act	bn:00053173n

Bibliographie

- Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A., and Guo, W. (2013). *SEM 2013 shared task : Semantic textual similarity. In Diab, M., Baldwin, T., and Baroni, M., editors, *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1 : Proceedings of the Main Conference and the Shared Task : Semantic Textual Similarity*, pages 32–43, Atlanta, Georgia, USA. Association for Computational Linguistics.
- Arora, H. S., Bhingardive, S., and Bhattacharyya, P. (2016). Detecting most frequent sense using word embeddings and BabelNet. In Fellbaum, C., Vossen, P., Mititelu, V. B., and Forascu, C., editors, *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 21–25, Bucharest, Romania. Global Wordnet Association.
- 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.
- 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.
- 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, Online. Association for Computational Linguistics.
- 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 Antoniadis, G., Blanchon, H., and Sérasset, G., editors, *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2 : TALN*, pages 321–334, Grenoble, France. ATALA/AFCP.
- Chaumartin, F.-R. (2007). WordNet et son écosystème : un ensemble de ressources linguistiques de large couverture. In *Colloque BD lexicales*, Montréal, Canada.

BIBLIOGRAPHIE

- Chen, Q., Zhu, X., Ling, Z.-H., Wei, S., Jiang, H., and Inkpen, D. (2017). Enhanced LSTM for natural language inference. In Barzilay, R. and Kan, M.-Y., editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 1657–1668, Vancouver, Canada. Association for Computational Linguistics.
- Ciaramita, M. and Johnson, M. (2003). Supersense tagging of unknown nouns in WordNet. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 168–175.
- Conia, S. and Navigli, R. (2020). Bridging the gap in multilingual semantic role labeling : a language-agnostic approach. In Scott, D., Bel, N., and Zong, C., editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1396–1410, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Conia, S. and Navigli, R. (2021). Framing word sense disambiguation as a multi-label problem for model-agnostic knowledge integration. 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 3269–3275, Online. Association for Computational Linguistics.
- Constant, M., Candito, M., and Seddah, D. (2013). The LIGM-Alpage architecture for the SPMRL 2013 shared task : Multiword expression analysis and dependency parsing. In Goldberg, Y., Marton, Y., Rehbein, I., and Versley, Y., editors, *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 46–52, Seattle, Washington, USA. 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.
- Dzikovska, M., Nielsen, R., Brew, C., Leacock, C., Giampiccolo, D., Bentivogli, L., Clark, P., Dagan, I., and Dang, H. T. (2013). SemEval-2013 task 7 : The joint student response analysis and 8th recognizing textual entailment challenge. 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 263–274, Atlanta, Georgia, USA. Association for Computational Linguistics.
- Eisenschlos, J., Ruder, S., Czapla, P., Kadras, M., Gugger, S., and Howard, J. (2019). MultiFiT : Efficient multi-lingual language model fine-tuning. 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 5702–5707, Hong Kong, China. Association for Computational Linguistics.
- Fellbaum, C. (1998). *WordNet : An Electronic Lexical Database*. The MIT Press.

- Flekova, L. and Gurevych, I. (2016). Supersense embeddings : A unified model for supersense interpretation, prediction, and utilization. In Erk, K. and Smith, N. A., editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 2029–2041, Berlin, Germany. Association for Computational Linguistics.
- Kitaev, N., Cao, S., and Klein, D. (2019). Multilingual constituency parsing with self-attention and pre-training. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505, Florence, Italy. Association for Computational Linguistics.
- Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., and Schwab, D. (2020). FlauBERT : des modèles de langue contextuels pré-entraînés pour le français (FlauBERT : Unsupervised language model pre-training for French). In Benoit, 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 268–278, Nancy, France. ATALA et AFCP.
- Liu, Z., Sun, Z., Cheng, S., Huang, S., and Wang, M. (2023). Only 5% attention is all you need : Efficient long-range document-level neural machine translation. In Park, J. C., Arase, Y., Hu, B., Lu, W., Wijaya, D., Purwarianti, A., and Krisnadhi, A. A., editors, *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 733–743, Nusa Dua, Bali. Association for Computational Linguistics.
- Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., Villemonte de la Clergerie, É., Sagot, B., and Seddah, D. (2020). Les modèles de langue contextuels camembert pour le français : impact de la taille et de l'hétérogénéité des données d'entraînement (CAMEMBERT contextual language models for French : Impact of training data size and heterogeneity). In Benoit, 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 54–65, Nancy, France. ATALA et AFCP.
- Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., and Joulin, A. (2018). Advances in pre-training distributed word representations. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Miller, G. A. (1994). WordNet : A lexical database for English. In *Human Language*

BIBLIOGRAPHIE

- Technology : Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994.*
- Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to WordNet : An On-line Lexical Database*. *International Journal of Lexicography*, 3(4) :235–244.
- 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.
- Nakov, P., Rosenthal, S., Kozareva, Z., Stoyanov, V., Ritter, A., and Wilson, T. (2013). SemEval-2013 task 2 : Sentiment analysis in Twitter. 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 312–320, Atlanta, Georgia, USA. Association for Computational Linguistics.
- Navigli, R. (2006). Meaningful clustering of senses helps boost word sense disambiguation performance. In Calzolari, N., Cardie, C., and Isabelle, P., editors, *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 105–112, Sydney, Australia. Association for Computational Linguistics.
- Navigli, R. (2009). Word sense disambiguation : A survey. *ACM Computing Surveys (CSUR)*, 41(2) :10.
- 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. 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, Georgia, USA. Association for Computational Linguistics.
- Navigli, R., Litkowski, K. C., and Hargraves, O. (2007). SemEval-2007 task 07 : Coarse-grained English all-words task. In Agirre, E., Màrquez, L., and Wicentowski, R., editors, *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 30–35, Prague, Czech Republic. Association for Computational Linguistics.
- Orlando, R., Conia, S., Brignone, F., Cecconi, F., and Navigli, R. (2021). AMuSE-WSD : An all-in-one multilingual system for easy Word Sense Disambiguation. In Adel, H. and Shi, S., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, pages 298–307, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- PALMER, M., DANG, H. T., and FELLBAUM, C. (2007). Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering*, 13(2) :137–163.

- Papandrea, S., Raganato, A., and Delli Bovi, C. (2017). SupWSD : A flexible toolkit for supervised word sense disambiguation. In Specia, L., Post, M., and Paul, M., editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, pages 103–108, Copenhagen, Denmark. Association for Computational Linguistics.
- Pedersen, B., Braasch, A., Johannsen, A., Alonso, H. M., Nimb, S., Olsen, S., Søgaard, A., and Sørensen, N. H. (2016). The SemDaX corpus — sense annotations with scalable sense inventories. 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 842–847, Portorož, Slovenia. European Language Resources Association (ELRA).
- Scarlino, B., Pasini, T., and Navigli, R. (2020). Sense-annotated corpora for word sense disambiguation in multiple languages and domains. 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 5905–5911, Marseille, France. European Language Resources Association.
- Schneider, N., Mohit, B., Oflazer, K., and Smith, N. A. (2012). Coarse lexical semantic annotation with supersenses : An Arabic case study. In Li, H., Lin, C.-Y., Osborne, M., Lee, G. G., and Park, J. C., editors, *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, pages 253–258, Jeju Island, Korea. 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, Online. Association for Computational Linguistics.
- 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, Gothenburg, Sweden. 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, Barcelona, Spain. Association for Computational Linguistics.
- Specia, L., Jauhar, S. K., and Mihalcea, R. (2012). SemEval-2012 task 1 : English lexical simplification. In Agirre, E., Bos, J., Diab, M., Manandhar, S., Marton, Y., and Yuret, D., editors, **SEM 2012 : The First Joint Conference on Lexical and Computational Semantics – Volume 1 : Proceedings of the main conference and the shared task, and Volume 2 : Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 347–355, Montréal, Canada. Association for Computational Linguistics.

BIBLIOGRAPHIE

- Stokoe, C., Oakes, M., and Tait, J. (2003). Word sense disambiguation in information retrieval revisited. page 159.
- Sumanth, C. and Inkpen, D. (2015). How much does word sense disambiguation help in sentiment analysis of micropost data ? In Balahur, A., van der Goot, E., Vossen, P., and Montoyo, A., editors, *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 115–121, Lisboa, Portugal. Association for Computational Linguistics.
- Vickrey, D., Biewald, L., Teyssier, M., and Koller, D. (2005). Word-sense disambiguation for machine translation. In Mooney, R., Brew, C., Chien, L.-F., and Kirchhoff, K., editors, *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 771–778, Vancouver, British Columbia, Canada. Association for Computational Linguistics.
- Wong, K., Paritosh, P., and Aroyo, L. (2021). Cross-replication reliability - an empirical approach to interpreting inter-rater reliability. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, pages 7053–7065, Online. Association for Computational Linguistics.