Désambiguïsation lexicale d'exemples lexicographiques du français : intégration de plongements de graphe et évaluation

Hee-Soo Choi[1, 2] Mathieu Constant[1] Karën Fort[2] Bruno Guillaume[2]
(1) ATILF, CNRS, Université de Lorraine, 54000 Nancy, France
(2) LORIA, Université de Lorraine, 54000 Nancy, France
utrucmuche@lab.fr, umachinchose@adresse-academique.be

RÉSUMÉ

Cet article présente une étude sur la désambiguïsation lexicale d'exemples lexicographique du français annotés avec des sens du RL-fr, un réseau lexico-sémantique de granularité fine. Nos expériences consistent à comparer différents types de systèmes, dont un modèle enrichi avec des plongements de graphe du RL-fr. Notre étude porte également une attention particulière à l'évaluation, avec l'introduction d'un nouveau cadre reflétant davantage les productions réelles des systèmes.

ABSTRACT

Word Sense Disambiguation of French Lexicographical Examples : integration of graph embeddings and evaluation.

This article presents a study on Word Sense Disambiguation of French lexicographical examples annotated with senses from RL-fr, a fine-grained lexical-semantic network. Our experiments consist in comparing different types of systems, including a model enriched with graph embeddings from RL-fr. Our study also pays particular attention to evaluation, with the introduction of a new evaluation framework that better reflects the actual outputs of the systems.

MOTS-CLÉS : désambiguïsation lexicale, évaluation, plongements de graphe.

KEYWORDS: word sense disambiguation, evaluation, graph embeddings.

1 Introduction

La désambiguïsation lexicale (*Word Sense Disambiguation*) est une tâche sémantique historique en Traitement Automatique des Langues (TAL) qui a fait l'objet d'un grand nombre de travaux ces dernières décennies (Bevilacqua *et al.*, 2021). La tâche consiste à prédire le sens correct d'un lexème polysémique, étant donné un contexte d'apparition. La majorité des données pour la désambiguisation lexicale est annotée avec des sens WordNet (Miller, 1995), tant pour l'anglais (Raganato *et al.*, 2017) que pour des données multilingues (Pasini *et al.*, 2021).

Dans cet article, nous nous intéressons à la désambiguïsation lexicale de données en français peu utilisées pour la tâche et annotées avec un inventaire de sens différent de WordNet : le corpus d'exemples lexicographiques du BEL-RL-fr (Lux-Pogodalla, 2014), annotés avec des sens du réseau lexico-sémantique RL-fr (Lux-Pogodalla & Polguère, 2011). Nos expériences s'inscrivent dans la continuité des travaux de Sinha *et al.* (2022) qui ont précédemment étudié l'utilisation du modèle EWISER (Bevilacqua & Navigli, 2020), intégrant des informations inhérentes au réseau

à travers une matrice de poids. Plus spécifiquement, nous explorons l'intégration de plongements de graphe du RL–fr dans EWISER. Dans un premier temps, nous évaluons différents types de modèles de désambiguïsation lexicale sur les noms communs et les verbes du BEL–RL–fr, allant de *baselines* produites grâce à des heuristiques à des grands modèles de langue auto-régressifs. Au-delà de comparer les performances des systèmes de désambiguïsation lexicale, notre recherche insiste sur le cadre d'évaluation et l'analyse qualitative des différents résultats. Notre article s'articule donc autour de deux questions de recherches : quels systèmes produisent les meilleures performances sur le corpus BEL–RL–fr dans un cadre d'évaluation classique ? (QR1) et quelle est l'influence de l'intégration des plongements de graphe dans le modèle EWISER ? (QR2).

2 BEL–RL–fr, un corpus annoté avec des sens du RL–fr

Le BEL–RL–fr (v2.0) (Lux-Pogodalla, 2014) est un corpus d'exemples lexicographiques issus de Frantext [1] et d'articles de l'Est Républicain dont certains lexèmes sont annotés avec des unités lexicales du RL–fr. Les mots annotés sont de différentes catégories grammaticales, telles que des noms communs, des verbes, des adjectifs, des adverbes ou des prépositions. L'inventaire de sens utilisé, le RL–fr (v2.1), est un réseau lexico-sémantique du français dont les nœuds décrivent des unités lexicales, et les arêtes, des relations lexico-sémantiques ou combinatoires. Les unités lexicales correspondent aux sens d'un **vocable**, représentant une entrée classique dans un dictionnaire. Par exemple, pour le lexème avocat, il existe deux vocables : les noms communs avocat^1 (métier) et avocat^2 (fruit). Chaque vocable peut admettre plusieurs sens, aussi appelés **acceptions** dans le RL–fr. Par exemple, avocat^1 (métier) admet deux sens : avocat^1_I, individu qui pratique le métier et avocat$^1_{II}$, individu qui a un certain comportement.

Le BEL–RL–fr a été initialement construit pour illustrer les contextes d'apparition des sens du RL–fr. Les exemples ayant été choisis manuellement, la fréquence d'apparition des sens n'est pas représentative de la distribution réelle dans un texte naturel, ce qui rend le corpus moins sensible au biais du sens le plus fréquent (Maru *et al.*, 2022). Pour nos expériences, nous nous concentrons sur les noms communs et les verbes, qui présentent le plus d'annotations en sens. Le tableau 1 présente les statistiques sur les annotations en sens pour ces catégories. On note que 71 % des sens n'apparaissent qu'une seule fois, avec des occurrences moyennes de 2,14 pour les noms et 1,90 pour les verbes.

	Noms	**Verbes**
Nombre d'exemples	20 783	8 631
Nombre d'instances annotées	28 587	9 950
Nombre de sens utilisés	13 359	5 237
Nombre de sens apparaissant une fois	6 463 (71 %)	3 715 (71 %)
Nombre de sens apparaissant deux fois	1 556 (12 %)	710 (14%)
Nombre de sens apparaissant trois fois	622 (5 %)	285 (5 %)
Nombre moyen d'occurrences d'un sens	2,14	1,90

TABLE 1 – Statistiques sur les exemples lexicographiques du BEL–RL–fr (v2).

Nous avons appliqué les mêmes pré-traitements et stratégie de division en échantillons d'entraînement, de validation et de test que Sinha *et al.* (2022). Les données sont filtrées pour ne garder que les exemples contenant des verbes ou des noms à désambiguïser. Par ailleurs, tous les exemples dans

1. https://www.frantext.fr/

lesquels un verbe ou un nom fait partie d'une expression polylexicale ont été retirés, ces exemples ne reflétant pas l'utilisation standard des mots qu'ils contiennent. Nous avons ensuite généré des plongements de sens sur `RL-fr` à partir d'une tâche de prédiction de liens. Au total, 18 965 exemples lexicographiques sont conservés pour les noms communs et 7 231 pour les verbes. Contrairement à certains jeux de données qui peuvent identifier plusieurs sens corrects pour un lexème-cible, dans `BEL-RL-fr`, chaque lexème est associé à un unique sens. Le niveau de polysémie, c-à-d le nombre moyen de sens possibles par lexème, est de 1,61 pour les noms et 2,12 pour les verbes. Après division, l'échantillon d'entraînement compte 13 645 instances pour les noms et 5 484 pour les verbes. Les échantillons de validation et de test contiennent 2 660 instances pour les noms et 874 pour les verbes.

3 Systèmes de désambiguïsation lexicale

Pour évaluer la désambiguïsation lexicale des exemples lexicographiques, nous reprenons le modèle `EWISER` de Sinha *et al.* (2022) auquel nous ajoutons deux nouvelles configurations : l'intégration d'une matrice de poids aléatoires et de poids calculés à partir des plongements de graphe. Nous couvrons également différents types de systèmes de désambiguïsation lexicale : (i) des *baselines* basées sur des heuristiques et une *baseline* neuronale, (ii) le modèle `EWISER` et ses quatre configurations et (iii) quatre grands modèles de langue auto-régressifs (`Mistral-7b`, `Qwen2.5-7b`, `Llama3.1-8b`, `Gemma3-4b`).

Nous appliquons trois *baselines* basées sur des heuristiques : une *baseline* aléatoire, la *baseline* classique du sens le plus fréquent (MFS) et une *baseline* basée sur le calcul de l'intersection entre le voisinage lexical des exemples et des sens dans le `RL-fr` avec un indice de Jaccard. Nous entraînons également un perceptron multicouche qui constitue une *baseline* neuronale supervisée (MLP).

Modèle `EWISER` et intégration de plongements de graphe. `EWISER` est un modèle neuronal supervisé pour la désambiguïsation lexicale enrichi par l'intégration d'informations provenant de `WordNet` (Bevilacqua & Navigli, 2020), possédant l'architecture suivante. Le plongement du lexème-cible $\boldsymbol{h}_t$, extrait via `CamemBERT`, est projeté dans l'espace des sens, normalisé par une couche BatchNorm (Luo *et al.*, 2019), puis transformé par la fonction d'activation $\text{swish}(x) = x \cdot \sigma(x)$ (Ramachandran *et al.*, 2018) :

$$\boldsymbol{u} = W(\text{BatchNorm}(\boldsymbol{h}_t)) + \boldsymbol{b}$$
$$\boldsymbol{h}' = \text{swish}(\boldsymbol{u})$$

où W et b sont des paramètres appris. Les probabilités finales de désambiguïsation sont obtenues en intégrant les connaissances structurelles du graphe de connaissances G via sa matrice d'adjacence $A \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$, où chaque entrée A_{s_1,s_2} correspond au poids de l'arc entre les synsets s_1 et s_2.

On définit d'abord les logits initiaux Z par une projection linéaire de la représentation cachée $\mathbf{h}'$: $Z = \mathbf{h}'O + b$, où $O \in \mathbb{R}^{d\times|\mathcal{V}|}$ est la matrice de poids et b le vecteur de biais. Pour permettre au modèle d'agréger les scores des synsets voisins dans le graphe, la matrice finale des logits Q est calculée selon : $Q = ZA^T + Z$. Les probabilités finales de désambiguïsation sont ensuite calculées via une fonction softmax appliquée à Q.

Contrairement à l'implémentation originale utilisant des plongements de sens externes (Bevilacqua & Navigli, 2020), nous suivons l'approche de Sinha *et al.* (2022) en nous concentrant sur l'intégration

de la connaissance structurelle via la matrice d'adjacence A (décrite précédemment) pour influencer ces prédictions.

Choix de A Sinha *et al.* (2022) proposent une adaptation du modèle pour la désambiguïsation du `BEL-RL-fr` avec les sens du `RL-fr`. Pour intégrer des informations provenant du graphe, `EWISER` utilise une matrice d'adjacence A permettant d'encoder des poids sur les relations entre les sens. Formellement, on considère un réseau lexico-sémantique : $G = \langle \mathcal{V}, \{R_{ij}, (i,j) \in \mathcal{V} \times \mathcal{V}\}, \omega \rangle$. $\mathcal{V}$ est l'ensemble des sens (nœuds). Pour deux sens i et j, R_{ij} est l'ensemble des relations entre i et j. Chacune de ces relations r possède un poids donné par une fonction $\omega(r) \in \mathbb{R}$. Nous pouvons construire une matrice d'adjacence A associée à G de la manière suivante : $A \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$, $\forall (i,j) \in \mathcal{V} \times \mathcal{V}$, $A_{ij} = \sum_{r \in R_{ij}} \omega(r)$.

Nous reprenons les deux stratégies de calcul des poids `STRUCT` et `SEM` proposées par Sinha *et al.* (2022) et nous ajoutons deux nouvelles stratégies, `COS` et `RANDOM` :

— `STRUCT` : on considère le **nombre de relations** entre deux sens.
$$\omega(r) = 1, \text{ donc } \forall (i,j) \in \mathcal{V} \times \mathcal{V}, A_{ij} = |R_{ij}|$$

— `SEM` : Une fonction lexicale (relation) entre deux unités lexicales présente un **poids sémantique** dans le `RL-fr`. Dans la stratégie `SEM`, le poids $\omega(r)$ est défini par :
$$\omega(r) = 1 + s_r, \text{ où } s_r \in \{0, 1, 2\} \text{ est le poids sémantique de la relation } r.$$

— `COS` : on considère la **similarité cosinus entre les plongements de graphe $\boldsymbol{\theta_i}, \boldsymbol{\theta_j}$** de deux sens i, j.
$$A_{ij} = cos_{sim}(\boldsymbol{\theta_i}, \boldsymbol{\theta_j})$$

— `RANDOM` : on fixe **aléatoirement** le poids entre deux sens.

Dans la stratégie `COS`, nous utilisons des plongements de graphe générés pour chaque sens du `RL-fr`. Nous testons l'hypothèse que le poids de la relation entre deux sens peut être représenté par la distance entre ses plongements de graphe respectifs. La stratégie `RANDOM` permet de comparer l'influence des valeurs des poids des différentes stratégies. Une fois la matrice A intégrée, il est possible de garder les poids « gelés » ou de les entraîner[2].

En *zero-shot*, la désambiguïsation lexicale avec les grands modèles de langue peut être abordée soit comme une tâche de classification, où le modèle sélectionne le sens correct parmi une liste, soit comme une tâche de génération, où il produit une définition du lexème-cible. La première approche permet une évaluation directe par comparaison au *gold*, tandis que la seconde, fondée sur la génération, rend l'évaluation plus complexe et repose généralement sur des métriques issues de la traduction automatique (Basile *et al.*, 2025; Meconi *et al.*, 2025). En l'absence de définitions dans le `RL-fr`, nous considérons donc la tâche comme une tâche de classification. Dans le *prompt*, les sens possibles sont décrits de deux manières (cf. Annexe A) : (i) par le voisinage lexical direct dans le réseau `RL-fr`, ce qui correspond à l'ensemble des sens liés au sens par une relation dans le réseau, toutes relations confondues (**lexical**), (ii) par un exemple d'utilisation du sens provenant du `BEL-RL-fr` (**usage**).

En raison de la forte sensibilité des modèles à la manière de formuler les *prompts* et à la langue (Mu *et al.*, 2024; Errica *et al.*, 2025; Shen *et al.*, 2024), nous testons tous les *prompts* en anglais et en français. Nous utilisons également un *prompt* demandant au modèle de générer une définition du

2. Les configurations avec la matrice entraînée seront notées avec un astérisque.

lexème à désambiguïser selon son contexte (**def**). Bien que la définition générée ne soit pas évaluée en l'absence de définition *gold*, cela nous permet d'observer si cette étape intermédiaire permet d'améliorer les performances du modèle. Nous obtenons ainsi un total de huit *prompts* différents. Pour les verbes, nous adoptons uniquement la stratégie lexicale, le `BEL-RL-fr` ne fournissant pas d'exemples lexicographiques pour tous les sens.

Quatre grands modèles de langue *open-weight* ont été testés, accessibles *via* la plate-forme Ollama[3] : `Gemma3-4b` (Team *et al.*, 2025), `Qwen2.5-7b` (Qwen *et al.*, 2025), `Llama3.1-8b` (Grattafiori *et al.*, 2024) et `Mistral-7b` (Jiang *et al.*, 2023). Pour des raisons de ressources matérielles disponibles, nous testons uniquement les versions avec le moins de paramètres.

4 Évaluer au-delà de la classification

La métrique traditionnelle pour évaluer la désambiguïsation lexicale est la micro-F1. Cette métrique est toutefois sensible à la fréquence d'apparition des sens. Un sens très fréquent et facilement prédit va augmenter la micro-F1, masquant ainsi la performance du système sur des sens peu fréquents. Nous évaluerons ainsi les performances des modèles avec une **micro-F1** et une **macro-F1**. Par ailleurs, afin d'analyser les prédictions d'un système supervisé de désambiguïsation lexicale, nous distinguons deux configurations d'évaluation :

— une évaluation **restreinte** : le sens prédit est le sens dont la probabilité est la plus haute **parmi les sens possibles du lexème-cible**. Cette configuration est celle traditionnellement appliquée dans les travaux sur la désambiguïsation lexicale.
— une évaluation **globale** : le sens prédit est le sens dont la probabilité est la plus haute **parmi tout le vocabulaire**.

La composante centrale d'un système supervisé de désambiguïsation lexicale est l'estimation de la probabilité conditionnelle : $P(s_i \mid c)$ où s_i est un sens du lexème-cible et c, le contexte d'apparition. Les systèmes neuronaux supervisés calculent un score pour chaque sens possible : $\text{score}(s_i, c) = f_\theta(s_i, c)$ où f_θ est une fonction neuronale paramétrée, basée sur des modèles de langue contextuels. Ces scores sont ensuite convertis en une distribution de probabilité avec une fonction softmax. Une fois la distribution de probabilités obtenue, le système choisit le sens avec la probabilité la plus haute.

En **évaluation restreinte**, la prédiction finale correspond au sens avec la probabilité la plus haute **parmi les sens possibles du lexème-cible**, $\mathcal{S}(w)$: $\hat{s}_{\mathcal{S}(w)} = \arg\max_{s_i \in \mathcal{S}(w)} P(s_i \mid c)$.

L'objectif étant d'évaluer si le système attribue le sens correct d'un lexème-cible parmi ses sens possibles, on se concentre uniquement sur les probabilités du sous-ensemble correspondant. En pratique, le système calcule la distribution de probabilités sur tout le vocabulaire, qui correspond à l'ensemble des sens possibles de tous les lexèmes. Il est alors possible de récupérer le sens avec la probabilité la plus haute **parmi tout le vocabulaire**, $\mathcal{V}$: $\hat{s}_{\mathcal{V}} = \arg\max_{s_i \in \mathcal{V}} P(s_i \mid c)$.

Cette configuration, que nous appellerons l'**évaluation globale**, permet d'observer si le système attribue une probabilité plus élevée à un sens qui n'est pas parmi les sens possibles. Si cette manière d'évaluer s'éloigne de l'objectif principal de la tâche de désambiguïsation lexicale, elle met davantage en valeur le comportement réel du modèle vis-à-vis des prédictions qu'il produit. Cette configuration n'est toutefois pas pertinente sur les *baselines* aléatoire et `MFS` et les grands modèles de langue

3. https://ollama.com/.

qui n'ont accès qu'aux sens possibles du lexème-cible. Par conséquent, l'évaluation restreinte et l'évaluation globale seront appliquées uniquement sur les modèles supervisés.

5 QR 1 : performances des systèmes en évaluation restreinte

Les modèles sont évalués sur le même jeu de données de test avec des performances données en macro-F1 (M-F1) et micro-F1 (m-F1) (cf. Tableau 2). Nous observons que les performances sont meilleures sur les noms que sur les verbes, ce qui est en accord avec les résultats de Sinha *et al.* (2022). Par ailleurs, les modèles les plus performants sont les mêmes quelle que soit la métrique considérée.

		Noms		Verbes	
Type de modèle	**Modèle**	**M-F1**	**m-F1**	**M-F1**	**m-F1**
Baselines heuristiques	`Random Sense`	13,23	23,95	6,59	14,77
	`MFS`	10,99	24,09	6,34	15,92
	`Voisinage lexical`	20,04	25,51	13,97	14,89
Baseline ML	`MLP`	47,35	56,09	23,03	31,96
`EWISER`	`EWISER-A`	**52,38**	**61,17**	30,28	40,78
	`EWISER+A RANDOM`	48,50	57,29	29,94	40,55
	`EWISER+A RANDOM*`	48,03	57,52	29,70	39,97
	`EWISER+A STRUCT`	49,57	57,29	33,16	41,47
	`EWISER+A STRUCT*`	51,07	58,46	31,38	40,32
	`EWISER+A SEM`	46,14	53,68	31,74	40,55
	`EWISER+A SEM*`	47,14	54,66	31,86	40,32
	`EWISER+A COS`	50,21	58,23	**33,27**	**43,76**
	`EWISER+A COS*`	50,86	58,76	32,74	43,30
`LLM`	`Qwen2.5-7b`	30,82	29,70	26,87	23,25
	`Gemma3-4b`	28,11	32,97	20,62	23,37
	`Llama3.1-8b`	28,00	33,23	23,42	19,82
	`Mistral-7b`	32,58	33,92	17,74	19,36

TABLE 2 – Performance (%) en macro-F1 (M-F1) et micro-F1 (m-F1) des modèles de désambiguïsation lexicale sur `BEL-RL-fr` sur les noms communs et les verbes en évaluation restreinte. Les meilleurs résultats sont présentés en gras.

Meilleures performances avec `EWISER`. Tant pour les noms que pour les verbes, le meilleur système est un modèle neuronal supervisé, `EWISER-A` pour les noms (m-F1 de 61,17 %) et `EWISER+A COS` pour les verbes (m-F1 de 43,76 %). L'ajout de la matrice A ne permet pas d'améliorer les performances sur les noms, quelle que soit la stratégie adoptée. Cependant, les performances sont meilleures avec `EWISER+A COS` pour les verbes, avec une m-F1 passant de 40,78 % à 43,30 %.

Concernant les *baselines*, la valeur de la *baseline* aléatoire est une moyenne de cinq tirages. Les résultats obtenus en m-F1 sont logiquement proches des exactitudes théoriques qui sont de 23,90 % pour les noms et 14,58 % pour les verbes. Nous remarquons ensuite que la *baseline* du sens le plus fréquent (`MFS`) est proche de la *baseline* aléatoire, la distribution des sens ne suivant pas une loi de Zipf comme dans les textes naturels. La *baseline* basée sur le voisinage lexical est également peu performante, avec des indices de Jaccard relativement faibles, l'indice maximal est de 0,25 pour les noms et 0,09 pour les verbes, ce qui montre une intersection limitée entre le voisinage lexical des sens dans le `RL-fr` et les lexèmes du contexte.

Le tableau 2 présente les meilleurs résultats en m-F1 pour chaque grand modèle de langue. Pour les noms, les résultats des modèles de langue sont plus ou moins constants à travers les modèles, avec une m-F1 moyenne de 32 %. Pour les verbes, on observe un écart un peu plus marqué entre `Qwen2.5-7b` et `Gemma3-4b` autour de 23 % de m-F1 et `Llama3.1-8b` et `Mistral-7b` autour de 19 %. Les performances sont plus élevées par rapport aux *baselines* heuristiques mais bien moindres qu'avec les modèles `EWISER`. Le meilleur modèle sur les noms est `Mistral-7b` avec un *prompt* en français décrivant les sens avec la stratégie **usage**. Sur les verbes, `Qwen2.5-7b` présente la M-F1 la plus élevée avec 26,87 % mais a une m-F1 de 23,25 % qui est légèrement plus basse que celle de `Gemma3-4b`, à 23,37 %.

Différences entre micro et macro-F1. Nous analysons l'écart de valeurs entre la m-F1 et la M-F1 pour les deux modèles les plus performants sur les noms communs et les verbes, respectivement `EWISER-A` et `EWISER+A COS`.

Pour les noms communs, `EWISER-A` présente une M-F1 de 52,38 % et une m-F1 de 61,17 %, ce qui représente un écart de 8,79 points. La m-F1 donne le même poids à chaque instance, contrairement à la M-F1 qui donne le même poids à chaque classe. En calculant la fréquence d'apparition des sens, c-à-d le nombre d'instances par sens dans le jeu de données de test, il existe 1 658 sens sur 2 660 instances pour les noms. On compte 63 % des sens n'apparaissant que dans une seule instance (1 044). Sur ces 1 044 instances, 54 % sont incorrectement prédites par le modèle, ce qui diminue fortement la M-F1. En revanche, ces sens mono-instance incorrectement prédits auront moins de poids dans un calcul de m-F1. Nous observons une tendance similaire sur les verbes, le modèle le plus performant, `EWISER+A COS`, présentant une M-F1 de 33,27 % et une m-F1 de 43,76 %, soit un écart de 10,49 points. Dans le jeu de données de test composé de 873 instances, 544 sens sont présents et 357 (66 %) d'entre eux n'apparaissent que dans une seule instance. Sur ces 357 instances, 71 % sont incorrectement prédites par le modèle.

6 QR 2 : influence de l'intégration de la matrice de poids

On se propose d'explorer l'influence de l'intégration de connaissances relatives au graphe *via* la matrice de poids A, dont les plongements de graphe du `RL-fr` avec la stratégie `COS`. Pour ce faire, nous observons plus spécifiquement les distributions de probabilités des prédictions.

Comparaison des prédictions des modèles. Nous avons observé que l'ajout de la matrice d'adjacence A fait baisser les performances d'`EWISER` pour les noms. Nous analysons une instance correctement désambiguïsée par `EWISER-A` mais incorrectement désambiguïsée par la majorité des modèles `EWISER+A` : « D'ailleurs les mamans non averties, qui cherchent en toute confiance des **chaussons** et un tablier de peinture pour une première rentrée en maternelle, après le 15 août, ne recommencent jamais cette erreur de parent d'élève débutant ! » (cit6158)

Le lexème **chausson** présente trois sens possibles dans le `RL-fr`[4] :

— `chausson I.1` : chaussures d'intérieur,
— `chausson I.2` : chaussures pour une activité donnée,
— `chausson II` : aliment.

4. https://spiderlex.atilf.fr/fr/q/*chausson***

Le sens correct dans l'instance en exemple est chausson I.1. Le tableau 3 présente le top 3 des sens s_i classés par leurs probabilités $P(s_i)$ des modèles EWISER.

Modèle	s_1	$P(s_1)$	s_2	$P(s_2)$	s_3	$P(s_3)$
EWISER-A	chausson I.1	0,5183	chausson I.2	0,4397	chausson II	0,0381
EWISER+A RANDOM	serviette III	0,5058	chausson I.2	0,3559	chausson I.1	0,1285
EWISER+A RANDOM*	chausson I.2	0,6846	chausson I.1	0,2953	chaussure I	0,0068
EWISER+A STRUCT	chausson I.2	0,8706	chausson I.1	0,1294	chausson II	0,0000
EWISER+A STRUCT*	chausson I.1	0,5571	chausson I.2	0,4429	chaussure I	0,0000
EWISER+A SEM	chausson I.2	0,9634	chausson I.1	0,0366	chausson II	0,0000
EWISER+A SEM*	chausson I.2	0,7941	chausson I.1	0,2059	chausson II	0,0000
EWISER+A COS	chausson I.2	0,9999	chausson I.1	0,0000	chaussonnier.N	0,0000
EWISER+A COS*	chausson I.2	0,9885	chausson I.1	0,0108	serviette III	0,0003

TABLE 3 – Top 3 des sens s_i et des probabilités $P(s_i)$ des modèles EWISER sur l'instance cit6158 ayant pour *gold* = chausson I.1.

Dans un premier temps, nous remarquons qu'EWISER-A prédit le sens correct avec une probabilité maximale de 0,5138. Le sens chausson I.2 est le deuxième sens au classement avec une probabilité de 0,4397 et enfin chausson II présente une probabilité plus faible à 0,0381. Les probabilités sont majoritairement distribuées sur les sens possibles du lexème-cible. Dans un second temps, sur les modèles intégrant la matrice A, six modèles sur huit prédisent incorrectement chausson I.2 avec une probabilité relativement élevée, allant de 0,6846 à 0,9999. Seul EWISER+A STRUCT* prédit correctement le sens correct avec une probabilité de 0,5571, qui est légèrement plus élevée qu'EWISER+A. La distinction entre chausson I.1 et chausson I.2 est sémantiquement fine, ce qui explique la confusion des modèles à prédire le sens correct. Par ailleurs, certains modèles EWISER+A prédisent des sens qui ne font pas partie des sens possibles du lexème-cible. Par exemple, EWISER+A RANDOM prédit serviette III en première position avec une probabilité de 0,5058 et le sens correct apparaît en troisième position avec une probabilité de 0,1285. Nous remarquons également la présence de sens tels que chaussure I ou chaussonnier.N. Nous supposons que les modèles intégrant la matrice sont influencés vers des sens reliés aux sens possibles du lexème-cible, se traduisant en une distribution des probabilités sur un ensemble de sens plus large que le sous-ensemble des sens possibles.

Pour les verbes, nous observons que le modèle EWISER+A COS présente les meilleures performances avec une micro-F1 de 43,76 % contre 40,78 % pour EWISER-A. Comme pour les noms, nous prenons l'exemple d'une instance correctement prédite par EWISER+A COS et nous comparons les prédictions des autres modèles : « Le savon **lave**-t-il mieux que le gel douche ? » (cit19262).

Le verbe **laver** présente dix sens différents dans le RL-fr. Dans cet exemple, le sens correct est laver II[5]. Le top 3 des sens prédits par chaque modèle EWISER et les probabilités correspondantes sont détaillés dans le tableau 4. EWISER+A COS prédit le sens correct avec une probabilité maximale de 0,4739, suivi du sens nettoyer I.1b avec une probabilité légèrement plus faible à 0,4282. Le sens correct n'est pas prédit avec un net écart de probabilités par rapport au second sens prédit, indiquant une proximité sémantique entre les deux sens, qui sont effectivement reliés par une relation de synonymie dans le RL-fr. Cela est confirmé par les résultats des autres modèles avec six sur huit modèles prédisant nettoyer I.1b avec des probabilités allant de 0,1904 à 0,9982.

Outre EWISER+A STRUCT et EWISER+A COS* qui prédisent le sens correct avec la probabilité maximale, nous remarquons que trois modèles prédisent laver II en deuxième position et deux

5. https://spiderlex.atilf.fr/fr/q/*laver***

autres modèles, en troisième position. Sur ces cinq derniers modèles, avec une évaluation classique, le sens correct est considéré comme bien prédit car celui-ci présente la probabilité maximale **parmi les sens possibles du lexème-cible**. Dans ce cadre d'évaluation, aucune différence n'est visible entre des sens prédits correctement avec une probabilité de 0,4739 (`EWISER+A COS`) et une probabilité de 0,0005 (`EWISER+A SEM*`). Par conséquent, une évaluation basée simplement sur une métrique telle qu'une m-F1 montre que sur cette instance, huit modèles sur neuf prédisent correctement le *gold* et un modèle se trompe. Afin de mettre en évidence ce phénomène, nous introduisons une nouvelle configuration d'évaluation, **l'évaluation globale**.

Modèle	s_1	$P(s_1)$	s_2	$P(s_2)$	s_3	$P(s_3)$
`EWISER+A COS`	`laver II`	0,4739	`nettoyer I.1b`	0,4282	`laver V.1`	0,0236
`EWISER-A`	`nettoyer I.1b`	0,2947	`laver II`	0,1933	`laver I.1`	0,1137
`EWISER+A RANDOM`	`nettoyer I.1b`	0,3238	`laver II`	0,1455	`laver VI`	0,1042
`EWISER+A RANDOM*`	`nettoyer I.1b`	0,1904	`laver VI`	0,1598	`laver I.1`	0,1009
`EWISER+A STRUCT`	`laver II`	0,2913	`nettoyer I.1b`	0,2912	`laver IV`	0,0959
`EWISER+A STRUCT*`	`nettoyer I.1b`	0,5556	`laver II`	0,1936	`laver VI`	0,0820
`EWISER+A SEM`	`nettoyer I.1b`	0,8777	`doubler V`	0,1032	`laver II`	0,0144
`EWISER+A SEM*`	`nettoyer I.1b`	0,9982	`bouffer I.1a`	0,0006	`laver II`	0,0005
`EWISER+A COS*`	`laver II`	0,4474	`nettoyer I.1b`	0,4300	`laver I.1`	0,0333

TABLE 4 – Top 3 des sens s et des probabilités des modèles `EWISER` sur l'instance cit19262 ayant pour *gold* = `laver II`.

Une baisse des scores en évaluation globale. Nous comparons les performances avec une évaluation globale, qui prend en compte le sens prédit à la probabilité maximale sur tout le vocabulaire, notée $\hat{s}_{\mathcal{V}}$. En évaluation restreinte, le sens prédit présente la probabilité maximale sur les sens possibles du lexème-cible, notée $\hat{s}_{\mathcal{S}(w)}$. On distingue ainsi deux cas de figure, décrits par des exemples dans le tableau 5 :

1. gold = $\hat{s}_{\mathcal{V}}$ $\wedge$ gold = $\hat{s}_{\mathcal{S}(w)}$: le sens *gold* est prédit correctement avec la probabilité la plus élevée sur tout le vocabulaire, donc sur les sens possibles du lexème-cible.
2. gold $\neq$ $\hat{s}_{\mathcal{V}}$ $\wedge$ gold = $\hat{s}_{\mathcal{S}(w)}$: le sens *gold* est prédit correctement mais ne présente pas la probabilité maximale sur tout le vocabulaire.

Le tableau 6 présente les performances en M-F1 et en m-F1 des modèles `EWISER`, sur les noms et sur les verbes, avec les deux types d'évaluation, globale et restreinte.

Cas	**Inst**	**gold**	$\hat{s}_{\mathcal{V}}$	P($\hat{s}_{\mathcal{V}}$)	$\hat{s}_{\mathcal{S}(w)}$	P($\hat{s}_{\mathcal{S}(w)}$)
1. gold = $\hat{s}_{\mathcal{V}}$ $\wedge$ gold = $\hat{s}_{\mathcal{S}(w)}$	cit16234	`témoin.N.masc II.2`	`témoin.N.masc II.2`	0,99	`témoin.N.masc II.2`	0,99
2. gold $\neq$ $\hat{s}_{\mathcal{V}}$ $\wedge$ gold = $\hat{s}_{\mathcal{S}(w)}$	cit29886	`terrain II`	`champ 1 II.1`	0,4807	`terrain II`	0,0113

TABLE 5 – Cas de figure des prédictions selon l'évaluation globale ou l'évaluation restreinte.

Le scénario optimal serait d'obtenir une m-F1 identique dans les deux évaluations, ce qui signifierait que tous les sens correctement prédits présentent la probabilité maximale sur tout le vocabulaire. Cependant, on remarque que la m-F1 en évaluation globale est plus faible qu'en évaluation restreinte. Il existe donc des instances où les modèles ont prédit le sens correct mais ont attribué la probabilité maximale à des sens qui ne font pas partie des sens possibles du lexème (cas de figure n°2). Pour les noms, l'écart entre la m-F1 globale et la m-F1 restreinte est plus marqué dans les

	Noms				Verbes			
	Évaluation globale		Évaluation restreinte		Évaluation globale		Évaluation restreinte	
Modèle	**M-F1**	**m-F1**	**M-F1**	**m-F1**	**M-F1**	**m-F1**	**M-F1**	**m-F1**
EWISER-A	**49,91**	**59,21**	**52,38**	**61,17**	27,69	38,03	30,28	40,78
EWISER+A RANDOM	44,83	54,29	48,50	57,29	27,76	37,91	29,94	40,55
EWISER+A RANDOM*	41,62	51,92	48,03	57,52	26,45	36,99	29,70	39,97
EWISER+A STRUCT	35,24	44,02	49,57	57,29	22,46	30,93	33,16	41,47
EWISER+A STRUCT*	42,57	51,32	51,07	58,46	23,59	33,22	31,38	40,32
EWISER+A SEM	31,44	40,23	46,14	53,68	22,57	30,81	31,74	40,55
EWISER+A SEM*	33,42	41,92	47,14	54,66	22,24	31,04	31,86	40,32
EWISER+A COS	43,91	52,86	50,21	58,23	28,35	39,29	**33,27**	**43,76**
EWISER+A COS*	44,49	53,53	50,86	58,76	**28,93**	**39,63**	32,74	43,30

TABLE 6 – Performances (%) en macro-F1 (M-F1) et micro-F1 (m-F1) des modèles EWISER sur les noms et les verbes, en évaluation globale et restreinte. Les meilleurs résultats sont donnés en gras.

modèles intégrant la matrice d'adjacence : en moyenne 8,22 points de différence (allant de 3 à 13,45 points) contre 1,96 points pour EWISER-A. Pour les verbes, les écarts les plus faibles sont de 2,64 pour EWISER+A RANDOM, 2,75 pour EWISER-A et 2,98 pour EWISER+A RANDOM*. Les autres modèles présentent des écarts compris entre 3,67 et 10,54.

Sur les noms, avec le modèle le plus performant, EWISER-A, il existe 52 instances où le sens correct est prédit mais ne présente pas la probabilité maximale. Ce nombre augmente sur les modèles qui intègrent la matrice A, notamment sur EWISER+A SEM, EWISER+A SEM*, et EWISER+A STRUCT avec plus de 300 instances. Ce phénomène explique la forte différence dans leur m-F1 restreinte et globale (-13 points en moyenne). Sur les verbes, les résultats sont différents dans la mesure où les deux modèles COS présentent les meilleures performances : 43,76 % pour EWISER+A COS et 43,30 % pour EWISER+A COS* contre 40,78 % pour EWISER-A en évaluation restreinte. Les modèles COS présentent davantage d'instances correctement prédites avec la probabilité maximale ($gold = \hat{s}_{\mathcal{V}}$) mais également d'instances correctement prédites sans la probabilité maximale ($gold \neq \hat{s}_{\mathcal{V}}$) par rapport à EWISER-A. Nous supposons que l'ajout de la matrice influence la distribution de probabilités vers des sens liés aux sens possibles du lexème-cible, ce qui expliquerait l'augmentation des instances $gold \neq \hat{s}_{\mathcal{V}}$.

7 Discussion et conclusion

Dans cette étude, nous nous sommes intéressés à la tâche de désambiguïsation lexicale sur des données peu utilisées pour la tâche, le corpus français BEL-RL-fr annoté avec des sens du RL-fr. À travers la QR 1, nous avons observé que le modèle supervisé EWISER était le plus performant sur les noms et sur les verbes par rapport aux *baselines* et aux grands modèles de langue. Concernant les performances générales des grands modèles de langue, nous avons observé qu'elles sont plus faibles que celles des modèles EWISER, indiquant des difficultés à désambiguïser en *zero-shot*. Cela peut être lié au fait que les sens ne sont pas décrits par des définitions, à la granularité fine des annotations du RL-fr et à l'utilisation des modèles de langues les plus petits en termes de paramètres.

Nous avons également exploré différentes configurations d'intégration de connaissances dans EWISER dans la QR 2. Nous avons observé que l'ajout de poids avec les plongements de graphe

entraîne de meilleures performances sur les verbes avec EWISER+A COS, mais pas sur les noms. Afin d'évaluer ce phénomène, nous avons introduit un nouveau cadre d'évaluation, appelée évaluation globale, qui nous a permis d'observer que l'ajout de la matrice de poids a une influence sur les distributions de probabilités. En effet, nous remarquons que les probabilités sont davantage distribuées sur des sens qui ne sont pas des sens possibles des lexèmes-cibles. Ce phénomène est davantage visible avec une évaluation globale où nous pouvons remarquer un écart plus élevé entre la micro-F1 globale et la micro-F1 restreinte dans les modèles EWISER+A.

Nous avons supposé que les sens qui ne font pas partie des sens possibles des lexèmes-cibles mais qui présentent des probabilités élevées, sont potentiellement des sens proches dans le graphe. Pour les noms, la distance moyenne entre les sens à la probabilité maximale et le *gold* est plus faible dans EWISER−A qu'avec les modèles EWISER+A, ce qui invalide notre hypothèse. Il semblerait que la matrice A génère du bruit plutôt que des informations utiles au modèle dans le cas des noms. Pour les verbes, le modèle EWISER+A COS présente de meilleures performances avec davantage d'instances correctement prédites. Si les probabilités moyennes des *gold* sont proches avec EWISER−A et EWISER+A COS, on observe que sur certaines instances, la distribution de probabilités est différente dans la mesure où EWISER+A COS peut attribuer des probabilités plus élevées à des sens autre que des sens possibles du lexème-cible, mais qui lui sont sémantiquement proches.

Bien que l'ajout d'une matrice de poids n'ait pas été concluante pour les noms, nous avons cherché à exploiter les plongements de graphe qui encodent des informations structurelles du réseau. Une limite de notre approche réside dans la modélisation de la proximité entre deux sens par une similarité cosinus entre leurs plongements de graphe respectifs, réduisant ainsi leur expressivité. Il serait alors intéressant d'intégrer ces connaissances, de la même manière que l'intégration externe de plongement de sens (Bevilacqua & Navigli, 2020).

Annexes

A Exemple de prompt

Références

Basile P., Siciliani L., Musacchio E. & Semeraro G. (2025). Exploring the Word Sense Disambiguation Capabilities of Large Language Models.

Bevilacqua M. & 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 D. Jurafsky, J. Chai, N. Schluter & J. Tetreault, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 2854–2864, Online : Association for Computational Linguistics. doi : 10.18653/v1/2020.acl-main.255.

Bevilacqua M., Pasini T., Raganato A. & Navigli R. (2021). Recent Trends in Word Sense Disambiguation : A Survey. In Z.-H. Zhou, Éd., *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, p. 4330–4338 : International Joint Conferences on Artificial Intelligence Organization. Survey Track, doi : 10.24963/ijcai.2021/593.

Errica F., Sanvito D., Siracusano G. & Bifulco R. (2025). What Did I Do Wrong ? Quantifying LLMs' Sensitivity and Consistency to Prompt Engineering. In L. Chiruzzo, A. Ritter

Prompt **Phrase d'entrée**	Étant donné le lexème cible dans la phrase d'entrée, choisis le sens correct parmi les candidats suivants. Génère uniquement le numéro de l'option sélectionnée. Le long du fleuve, en été, toute sa famille se promène en `bicyclette`.
Description 1 : `RL-fr` voisinage	Les candidats sont : 1) ['cycliste', 'véhicule roulant', 'enfourcher', 'bécane', 'circuler', 'cyclotouriste', 'biclo', 'deux-roues', 'à', 'déplacer', 'guidon', 'en', 'cycle', 'descendre', 'monter', 'avoir l'air d'un crapaud sur une boîte d'allumettes', 'petite reine', 'rouler', 'bicyclette', 'escalader', 'escalade', 'grimper', 'pédaler', 'clou', 'vélocipède', 'biclou', 'vélo', 'faire', 'cyclotouriste'] 2) ['bicyclette', 'faire', 'cyclisme', 'vélo', 'cycliste', 'pratiquer', 'petite reine', 'loisir', 'cyclable']
Description 2 : `BEL-RL-fr` exemples	Les candidats sont : 1) La bicyclette était à ma taille, elle avait une lampe ronde et une dynamo, j'en caressais sans cesse le volume, la consistance, et je n'arrivais pas à croire qu'elle était à moi, que je pouvais avoir de tels objets « en vrai ». 2) L'agglomération de Nancy lance à son tour un système de vélos en libre service. Petite innovation par rapport aux initiatives précédentes en la matière : une Maison du vélo sera prochainement ouverte afin de promouvoir la pratique de la bicyclette

TABLE 7 – Deux types de description des sens possibles du lexème `bicyclette`.

& L. WANG, Éds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics : Human Language Technologies (Volume 1 : Long Papers)*, p. 1543–1558, Albuquerque, États-Unis : Association for Computational Linguistics. DOI : 10.18653/v1/2025.naacl-long.73.

GRATTAFIORI A., DUBEY A., JAUHRI A., PANDEY A., KADIAN A., AL-DAHLE A., LETMAN A., MATHUR A., SCHELTEN A., VAUGHAN A., YANG A., FAN A., GOYAL A., HARTSHORN A., YANG A., MITRA A., SRAVANKUMAR A., KORENEV A., HINSVARK A., RAO A., ZHANG A., RODRIGUEZ A., GREGERSON A., SPATARU A., ROZIERE B., BIRON B., TANG B., CHERN B., CAUCHETEUX C., NAYAK C., BI C., MARRA C., MCCONNELL C., KELLER C., TOURET C., WU C., WONG C., FERRER C. C., NIKOLAIDIS C., ALLONSIUS D., SONG D., PINTZ D., LIVSHITS D., WYATT D., ESIOBU D., CHOUDHARY D., MAHAJAN D., GARCIA-OLANO D., PERINO D., HUPKES D., LAKOMKIN E., ALBADAWY E., LOBANOVA E., DINAN E., SMITH E. M., RADENOVIC F., GUZMÁN F., ZHANG F., SYNNAEVE G., LEE G., ANDERSON G. L., THATTAI G., NAIL G., MIALON G., PANG G., CUCURELL G., NGUYEN H., KOREVAAR H., XU H., TOUVRON H., ZAROV I., IBARRA I. A., KLOUMANN I., MISRA I., EVTIMOV I., ZHANG J., COPET J., LEE J., GEFFERT J., VRANES J., PARK J., MAHADEOKAR J., SHAH J., VAN DER LINDE J., BILLOCK J., HONG J., LEE J., FU J., CHI J., HUANG J., LIU J., WANG J., YU J., BITTON J., SPISAK J., PARK J., ROCCA J., JOHNSTUN J., SAXE J., JIA J., ALWALA K. V., PRASAD K., UPASANI K., PLAWIAK K., LI K., HEAFIELD K., STONE K., EL-ARINI K., IYER K., MALIK K., CHIU K., BHALLA K., LAKHOTIA K., RANTALA-YEARY L., VAN DER MAATEN L., CHEN L., TAN L., JENKINS L., MARTIN L., MADAAN L., MALO L., BLECHER L., LANDZAAT L., DE OLIVEIRA L., MUZZI M., PASUPULETI M., SINGH M., PALURI M., KARDAS M., TSIMPOUKELLI M., OLDHAM M., RITA M., PAVLOVA M., KAMBADUR M., LEWIS M., SI M., SINGH M. K., HASSAN M., GOYAL N., TORABI N., BASHLYKOV N., BOGOYCHEV N.,

Chatterji N., Zhang N., Duchenne O., Çelebi O., Alrassy P., Zhang P., Li P., Vasic P., Weng P., Bhargava P., Dubal P., Krishnan P., Koura P. S., Xu P., He Q., Dong Q., Srinivasan R., Ganapathy R., Calderer R., Cabral R. S., Stojnic R., Raileanu R., Maheswari R., Girdhar R., Patel R., Sauvestre R., Polidoro R., Sumbaly R., Taylor R., Silva R., Hou R., Wang R., Hosseini S., Chennabasappa S., Singh S., Bell S., Kim S. S., Edunov S., Nie S., Narang S., Raparthy S., Shen S., Wan S., Bhosale S., Zhang S., Vandenhende S., Batra S., Whitman S., Sootla S., Collot S., Gururangan S., Borodinsky S., Herman T., Fowler T., Sheasha T., Georgiou T., Scialom T., Speckbacher T., Mihaylov T., Xiao T., Karn U., Goswami V., Gupta V., Ramanathan V., Kerkez V., Gonguet V., Do V., Vogeti V., Albiero V., Petrovic V., Chu W., Xiong W., Fu W., Meers W., Martinet X., Wang X., Wang X., Tan X. E., Xia X., Xie X., Jia X., Wang X., Goldschlag Y., Gaur Y., Babaei Y., Wen Y., Song Y., Zhang Y., Li Y., Mao Y., Coudert Z. D., Yan Z., Chen Z., Papakipos Z., Singh A., Srivastava A., Jain A., Kelsey A., Shajnfeld A., Gangidi A., Victoria A., Goldstand A., Menon A., Sharma A., Boesenberg A., Baevski A., Feinstein A., Kallet A., Sangani A., Teo A., Yunus A., Lupu A., Alvarado A., Caples A., Gu A., Ho A., Poulton A., Ryan A., Ramchandani A., Dong A., Franco A., Goyal A., Saraf A., Chowdhury A., Gabriel A., Bharambe A., Eisenman A., Yazdan A., James B., Maurer B., Leonhardi B., Huang B., Loyd B., Paola B. D., Paranjape B., Liu B., Wu B., Ni B., Hancock B., Wasti B., Spence B., Stojkovic B., Gamido B., Montalvo B., Parker C., Burton C., Mejia C., Liu C., Wang C., Kim C., Zhou C., Hu C., Chu C.-H., Cai C., Tindal C., Feichtenhofer C., Gao C., Civin D., Beaty D., Kreymer D., Li D., Adkins D., Xu D., Testuggine D., David D., Parikh D., Liskovich D., Foss D., Wang D., Le D., Holland D., Dowling E., Jamil E., Montgomery E., Presani E., Hahn E., Wood E., Le E.-T., Brinkman E., Arcaute E., Dunbar E., Smothers E., Sun F., Kreuk F., Tian F., Kokkinos F., Ozgenel F., Caggioni F., Kanayet F., Seide F., Florez G. M., Schwarz G., Badeer G., Swee G., Halpern G., Herman G., Sizov G., Guangyi, Zhang, Lakshminarayanan G., Inan H., Shojanazeri H., Zou H., Wang H., Zha H., Habeeb H., Rudolph H., Suk H., Aspegren H., Goldman H., Zhan H., Damlaj I., Molybog I., Tufanov I., Leontiadis I., Veliche I.-E., Gat I., Weissman J., Geboski J., Kohli J., Lam J., Asher J., Gaya J.-B., Marcus J., Tang J., Chan J., Zhen J., Reizenstein J., Teboul J., Zhong J., Jin J., Yang J., Cummings J., Carvill J., Shepard J., McPhie J., Torres J., Ginsburg J., Wang J., Wu K., U K. H., Saxena K., Khandelwal K., Zand K., Matosich K., Veeraraghavan K., Michelena K., Li K., Jagadeesh K., Huang K., Chawla K., Huang K., Chen L., Garg L., A L., Silva L., Bell L., Zhang L., Guo L., Yu L., Moshkovich L., Wehrstedt L., Khabsa M., Avalani M., Bhatt M., Mankus M., Hasson M., Lennie M., Reso M., Groshev M., Naumov M., Lathi M., Keneally M., Liu M., Seltzer M. L., Valko M., Restrepo M., Patel M., Vyatskov M., Samvelyan M., Clark M., Macey M., Wang M., Hermoso M. J., Metanat M., Rastegari M., Bansal M., Santhanam N., Parks N., White N., Bawa N., Singhal N., Egebo N., Usunier N., Mehta N., Laptev N. P., Dong N., Cheng N., Chernoguz O., Hart O., Salpekar O., Kalinli O., Kent P., Parekh P., Saab P., Balaji P., Rittner P., Bontrager P., Roux P., Dollar P., Zvyagina P., Ratanchandani P., Yuvraj P., Liang Q., Alao R., Rodriguez R., Ayub R., Murthy R., Nayani R., Mitra R., Parthasarathy R., Li R., Hogan R., Battey R., Wang R., Howes R., Rinott R., Mehta S., Siby S., Bondu S. J., Datta S., Chugh S., Hunt S., Dhillon S., Sidorov S., Pan S., Mahajan S., Verma S., Yamamoto S., Ramaswamy S., Lindsay S., Lindsay S., Feng S., Lin S., Zha S. C., Patil

S., Shankar S., Zhang S., Zhang S., Wang S., Agarwal S., Sajuyigbe S., Chintala S., Max S., Chen S., Kehoe S., Satterfield S., Govindaprasad S., Gupta S., Deng S., Cho S., Virk S., Subramanian S., Choudhury S., Goldman S., Remez T., Glaser T., Best T., Koehler T., Robinson T., Li T., Zhang T., Matthews T., Chou T., Shaked T., Vontimitta V., Ajayi V., Montanez V., Mohan V., Kumar V. S., Mangla V., Ionescu V., Poenaru V., Mihailescu V. T., Ivanov V., Li W., Wang W., Jiang W., Bouaziz W., Constable W., Tang X., Wu X., Wang X., Wu X., Gao X., Kleinman Y., Chen Y., Hu Y., Jia Y., Qi Y., Li Y., Zhang Y., Zhang Y., Adi Y., Nam Y., Yu, Wang, Zhao Y., Hao Y., Qian Y., Li Y., He Y., Rait Z., DeVito Z., Rosnbrick Z., Wen Z., Yang Z., Zhao Z. & Ma Z. (2024). The Llama 3 Herd of Models.

Jiang A. Q., Sablayrolles A., Mensch A., Bamford C., Chaplot D. S., de las Casas D., Bressand F., Lengyel G., Lample G., Saulnier L., Lavaud L. R., Lachaux M.-A., Stock P., Scao T. L., Lavril T., Wang T., Lacroix T. & Sayed W. E. (2023). Mistral 7B.

Luo P., Wang X., Shao W. & Peng Z. (2019). Towards Understanding Regularization in Batch Normalization. In *International Conference on Learning Representations*, La Nouvelle-Orléans, États-Unis.

Lux-Pogodalla V. (2014). Integrating lexicographic examples in a lexical network (Intégration relationnelle des exemples lexicographiques dans un réseau lexical) [in French]. In *Proceedings of TALN 2014 (Volume 2 : Short Papers)*, p. 586–591, Marseille, France : ATALA.

Lux-Pogodalla V. & Polguère A. (2011). Construction of a French Lexical Network : Methodological Issues. In *First InternationalWorkshop on Lexical Resources, WoLeR 2011*, p. 54–61, Ljubljana, Slovénie. HAL : hal-00686467.

Maru M., Conia S., Bevilacqua M. & Navigli R. (2022). Nibbling at the Hard Core of Word Sense Disambiguation. In S. Muresan, P. Nakov & A. Villavicencio, Éds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 4724–4737, Dublin, Irlande : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.324.

Meconi D., Stirpe S., Martelli F., Lavalle L. & Navigli R. (2025). Do Large Language Models Understand Word Senses ? In C. Christodoulopoulos, T. Chakraborty, C. Rose & V. Peng, Éds., *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, p. 33885–33904, Suzhou, Chine : Association for Computational Linguistics. DOI : 10.18653/v1/2025.emnlp-main.1720.

Miller G. A. (1995). WordNet : A Lexical Database for English. *Communications of the ACM*, **38**(11), 39–41.

Mu Y., Wu B. P., Thorne W., Robinson A., Aletras N., Scarton C., Bontcheva K. & Song X. (2024). Navigating Prompt Complexity for Zero-Shot Classification : A Study of Large Language Models in Computational Social Science. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti & N. Xue, Éds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, p. 12074–12086, Turin, Italie : ELRA and ICCL.

Pasini T., Raganato A. & Navigli R. (2021). XL-WSD : An Extra-Large and Cross-Lingual Evaluation Framework for Word Sense Disambiguation. *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**(15), 13648–13656. DOI : 10.1609/aaai.v35i15.17609.

Qwen, :, Yang A., Yang B., Zhang B., Hui B., Zheng B., Yu B., Li C., Liu D., Huang F., Wei H., Lin H., Yang J., Tu J., Zhang J., Yang J., Yang J., Zhou J., Lin J., Dang K., Lu K., Bao K., Yang K., Yu L., Li M., Xue M., Zhang P., Zhu Q., Men R., Lin R.,

LI T., TANG T., XIA T., REN X., REN X., FAN Y., SU Y., ZHANG Y., WAN Y., LIU Y., CUI Z., ZHANG Z. & QIU Z. (2025). Qwen2.5 Technical Report.

RAGANATO A., CAMACHO-COLLADOS J. & NAVIGLI R. (2017). Word Sense Disambiguation : A Unified Evaluation Framework and Empirical Comparison. In M. LAPATA, P. BLUNSOM & A. KOLLER, Éds., *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 1, Long Papers*, p. 99–110, Valence, Espagne : Association for Computational Linguistics.

RAMACHANDRAN P., ZOPH B. & LE Q. V. (2018). Searching for Activation Functions. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, Vancouver, Canada. Workshop Track.

SHEN L., TAN W., CHEN S., CHEN Y., ZHANG J., XU H., ZHENG B., KOEHN P. & KHASHABI D. (2024). The Language Barrier : Dissecting Safety Challenges of LLMs in Multilingual Contexts. In L.-W. KU, A. MARTINS & V. SRIKUMAR, Éds., *Findings of the Association for Computational Linguistics : ACL 2024*, p. 2668–2680, Bangkok, Thaïlande : Association for Computational Linguistics. DOI : 10.18653/v1/2024.findings-acl.156.

SINHA A., OLLINGER S. & CONSTANT M. (2022). Word Sense Disambiguation of French Lexicographical Examples Using Lexical Networks. In D. USTALOV, Y. GAO, A. PANCHENKO, M. VALENTINO, M. THAYAPARAN, T. H. NGUYEN, G. PENN, A. RAMESH & A. JANA, Éds., *Proceedings of TextGraphs-16 : Graph-based Methods for Natural Language Processing*, p. 70–76, Gyeongju, République de Corée : Association for Computational Linguistics.

TEAM G., KAMATH A., FERRET J., PATHAK S., VIEILLARD N., MERHEJ R., PERRIN S., MATEJOVICOVA T., RAMÉ A., RIVIÈRE M., ROUILLARD L., MESNARD T., CIDERON G., BASTIEN GRILL J., RAMOS S., YVINEC E., CASBON M., POT E., PENCHEV I., LIU G., VISIN F., KENEALY K., BEYER L., ZHAI X., TSITSULIN A., BUSA-FEKETE R., FENG A., SACHDEVA N., COLEMAN B., GAO Y., MUSTAFA B., BARR I., PARISOTTO E., TIAN D., EYAL M., CHERRY C., PETER J.-T., SINOPALNIKOV D., BHUPATIRAJU S., AGARWAL R., KAZEMI M., MALKIN D., KUMAR R., VILAR D., BRUSILOVSKY I., LUO J., STEINER A., FRIESEN A., SHARMA A., SHARMA A., GILADY A. M., GOEDECKEMEYER A., SAADE A., FENG A., KOLESNIKOV A., BENDEBURY A., ABDAGIC A., VADI A., GYÖRGY A., PINTO A. S., DAS A., BAPNA A., MIECH A., YANG A., PATERSON A., SHENOY A., CHAKRABARTI A., PIOT B., WU B., SHAHRIARI B., PETRINI B., CHEN C., LAN C. L., CHOQUETTE-CHOO C. A., CAREY C., BRICK C., DEUTSCH D., EISENBUD D., CATTLE D., CHENG D., PAPARAS D., SREEPATHIHALLI D. S., REID D., TRAN D., ZELLE D., NOLAND E., HUIZENGA E., KHARITONOV E., LIU F., AMIRKHANYAN G., CAMERON G., HASHEMI H., KLIMCZAK-PLUCIŃSKA H., SINGH H., MEHTA H., LEHRI H. T., HAZIMEH H., BALLANTYNE I., SZPEKTOR I., NARDINI I., POUGET-ABADIE J., CHAN J., STANTON J., WIETING J., LAI J., ORBAY J., FERNANDEZ J., NEWLAN J., YEONG JI J., SINGH J., BLACK K., YU K., HUI K., VODRAHALLI K., GREFF K., QIU L., VALENTINE M., COELHO M., RITTER M., HOFFMAN M., WATSON M., CHATURVEDI M., MOYNIHAN M., MA M., BABAR N., NOY N., BYRD N., ROY N., MOMCHEV N., CHAUHAN N., SACHDEVA N., BUNYAN O., BOTARDA P., CARON P., RUBENSTEIN P. K., CULLITON P., SCHMID P., SESSA P. G., XU P., STANCZYK P., TAFTI P., SHIVANNA R., WU R., PAN R., ROKNI R., WILLOUGHBY R., VALLU R., MULLINS R., JEROME S., SMOOT S., GIRGIN S., IQBAL S., REDDY S., SHETH S., PÕDER S., BHATNAGAR S., PANYAM S. R., EIGER S., ZHANG S., LIU T., YACOVONE T., LIECHTY T., KALRA U., EVCI U., MISRA V., ROSEBERRY V., FEINBERG V., KOLESNIKOV V., HAN W., KWON W., CHEN X., CHOW Y., ZHU Y., WEI Z., EGYED Z., COTRUTA V., GIANG M., KIRK P., RAO A., BLACK K., BABAR N., LO J., MOREIRA E., MARTINS L. G., SANSEVIERO O., GONZALEZ L., GLEICHER Z., WARKENTIN T., MIRROKNI V., SENTER E., COLLINS E.,

BARRAL J., GHAHRAMANI Z., HADSELL R., MATIAS Y., SCULLEY D., PETROV S., FIEDEL N., SHAZEER N., VINYALS O., DEAN J., HASSABIS D., KAVUKCUOGLU K., FARABET C., BUCHATSKAYA E., ALAYRAC J.-B., ANIL R., DMITRY, LEPIKHIN, BORGEAUD S., BACHEM O., JOULIN A., ANDREEV A., HARDIN C., DADASHI R. & HUSSENOT L. (2025). Gemma 3 Technical Report.