



DÉSAMBIGUISATION AUTOMATIQUE D'UN TERME POLYSÉMIQUE

Le cas du mot femme

Mathilde PIERRON

Sous la direction de Cécile FABRE et Ludovic TANGUY



2025-2026

DÉPARTEMENT DES SCIENCES DU LANGAGE
MASTER 1 – Linguistique, Informatique et Technologies du Langage

Table des matières

Introduction	1
Partie 1. La polysémie : définitions et impacts sur le Traitement Automatique des Langues (TAL).....	3
1.1. La polysémie : concepts-clés	3
1.1.1. Caractéristiques de la polysémie	3
1.1.2. Ambiguïté, homonymie et monosémie.....	4
1.1.3. Les différents types de polysémie	5
1.1.4. Les limites et les débats autour de cette notion	5
1.1.5. La représentation des sens dans les dictionnaires	6
1.1.6. Traitement cognitif des sens multiples	6
1.2. Le traitement automatique de la polysémie.....	7
1.2.1. Les ressources pour la désambiguïsation	7
1.2.2. La notion de sens en TAL	8
1.2.3. Approche symbolique : l'algorithme de Lesk	9
1.2.4. Représentations vectorielles des mots	10
1.2.5. Les approches principales de la WSD	10
1.2.6. Approches basées sur la traduction	11
1.2.7. Évaluation des performances	12
1.2.8. La place de la désambiguïsation lexicale dans le contexte des LLM.....	13
1.3. Le mot <i>femme</i> : évolution sémantique	14
Partie 2. Analyse des régularités d'emploi du mot <i>femme</i>	17
2.1. La polysémie du mot <i>femme</i> dans les dictionnaires	17
2.2. <i>Frantext</i> : recherche de patrons morphosyntaxiques	19
2.3. <i>Word Sketch Difference</i> : <i>femme</i> et <i>épouse</i>	21
2.4. Synthèse des régularités observées.....	23
Partie 3. Élaboration d'un système automatique de désambiguïsation	25
3.1. Méthodologie de conception du système automatique	25
3.2. Formalisation des règles linguistiques et difficultés identifiées	26
3.3. Annotation des cas limites.....	29
3.4. Fréquence d'apparition des règles dans le système final	31
Partie 4. Évaluation du système de désambiguïsation.....	33
4.1. Constitution d'un <i>gold standard</i> à partir du corpus <i>OPUS</i>	33
4.1.1. Choix d'un corpus parallèle anglais-français	33
4.1.2. Les différentes traductions de <i>femme</i> en anglais.....	35
4.1.3. Méthodologie de constitution du <i>gold standard</i>	36
4.2. Limites du corpus <i>OPUS</i>	37
4.3. Évaluation des performances	38

4.4. Comparaison des performances avec un LLM	39
<i>Partie 5. Variations inter-corpus et évolution diachronique des usages de femme</i>	43
5.1. Variations des usages de <i>femme</i> selon les corpus	44
5.1.1. Présentation des corpus comparés.....	44
5.1.2. Préparation des échantillons pour l'annotation automatique	45
5.1.3. Vérification des performances du système selon les corpus	46
5.1.4. Répartition des deux sens de <i>femme</i> selon les corpus	47
5.1.5. Variation des règles de désambiguïsation selon les corpus.....	49
5.2. Évolution diachronique des usages de <i>femme</i> et d' <i>épouse</i> dans <i>Frantext</i>	50
5.2.1. Présentation du corpus et des périodes étudiées	50
5.2.2. Résultats de l'analyse diachronique.....	51
<i>Conclusion et perspectives.....</i>	53
Bilan	53
Limites.....	54
Perspectives	55
<i>Bibliographie</i>	56
<i>Annexes</i>	59

Remerciements

Je tiens à adresser mes remerciements à mes deux encadrants de mémoire, Ludovic Tanguy et Cécile Fabre, pour la qualité de leurs conseils, leur disponibilité et l'attention qu'ils ont portée à mon travail tout au long de cette année. Leur accompagnement m'a été précieux dans la réalisation de ce mémoire.

Je souhaite remercier les trois enseignants du master LITL, Lydia Mai Ho-Dac, Ludovic Tanguy et Cécile Fabre, pour leur pédagogie dynamique et accessible, ainsi que pour leur investissement auprès des étudiants. Le master LITL n'aurait pas eu la même richesse sans leur présence et leur engagement.

Je remercie particulièrement Jérémy, qui m'a aidée à trouver ma voie et sans qui je n'aurais probablement pas eu le courage de reprendre mes études. Je n'oublierai pas le soutien qu'il m'apporte depuis dix ans, ni ses encouragements et les sacrifices qu'il a faits pour me permettre de mener à bien ces années d'études.

Merci à mes trois annotateurs, Marilou, Hugo et Jean-Baptiste, qui m'ont aidée à prendre certaines décisions d'annotation. Je remercie une seconde fois Jean-Baptiste pour ses relectures, pour avoir accepté de servir de testeur lorsque j'avais des doutes sur certains usages, ainsi que pour son soutien indéfectible dans nos projets.

Mes remerciements vont également à toute la promotion LITL 2025-2026, avec qui j'ai passé une formidable année. Ce fut très enrichissant de travailler à vos côtés. Un grand merci en particulier à Paula, ma partenaire d'emploi du temps cette année ; à Herandi, pour sa capacité à voir chaque travail comme une occasion de développer de nouvelles compétences ; et à Marilou et Hugo, avec qui j'ai souvent travaillé sur les devoirs.

Enfin, je tiens à remercier tous mes proches, ainsi que toutes les personnes qui ont contribué, d'une manière ou d'une autre, à la réussite de cette année.

Introduction

La polysémie, c'est-à-dire la capacité d'un mot à posséder plusieurs sens reliés sémantiquement pour une même forme lexicale, occupe une place centrale dans le fonctionnement des langues naturelles (Mazaleyrat, 2010). En effet, en anglais par exemple, environ 40 % des mots fréquents seraient polysémiques, faisant de la polysémie une propriété largement répandue dans le lexique plutôt qu'un phénomène rare (Durkin et Manning, 1989, cités dans Haber et Poesio, 2024). Cette pluralité de sens constitue une source d'ambiguïté qui complique le travail d'analyse, notamment lorsque l'on cherche à automatiser l'interprétation du lexique. Le vocabulaire de la famille permet d'aborder cette question à partir d'un domaine lexical à première vue très organisé. Les termes de parenté se définissent en effet les uns par rapport aux autres, mais ils peuvent aussi recevoir des interprétations différentes selon les contextes dans lesquels ils sont employés.

C'est dans ce cadre que ce mémoire prend pour point de départ les travaux de Tanguy et al. (2024, 2025), portant sur l'analyse distributionnelle du vocabulaire de la famille. Leur première étude (Tanguy et al., 2024) vise à analyser comment les termes de parenté (*fils*, *tante*, *neveu*, *grand-mère*, etc.) s'organisent sémantiquement dans différents corpus français (plusieurs genres textuels : *Wikipédia*, *revues.org*, *Frantext*, *Le Monde*, *FrWaC (French Web as Corpus)*) à l'aide de plongements de mots¹ (*embeddings*), afin de voir si ces modèles captent des traits structurants comme le genre, l'ascendance ou l'alliance. La seconde étude (Tanguy et al., 2025) compare trois méthodes : le plongement de mots, le *topic modeling*² à l'aide du corpus *Wikipédia* contenant 256 millions de mots et le modèle de langue (*Large Language Model* ou LLM³) *ChatGPT*. Cela leur a permis d'analyser le vocabulaire de la famille, afin d'évaluer la capacité de ces techniques à représenter à la fois les traits structurels de la parenté et certaines thématiques associées.

Dans leurs recherches, ces chercheurs ont rencontré certaines limites avec les termes polysémiques, en particulier avec le terme *femme*, qu'ils ont dû exclure de leur étude pour éviter les résultats biaisés. En effet, leur hypothèse est que ce terme serait majoritairement présent dans les corpus considérés au sens *d'être humain de sexe féminin*, et non dans le sens de la famille (*épouse*), qui les intéresse pour leur étude. Cette forte prédominance d'un sens non pertinent dans le contexte de leur étude le distingue d'autres termes polysémiques, pour lesquels les différents usages restent mieux répartis. En effet, ils ont choisi de garder le terme *filles*, qui est utilisé dans ses deux sens (*par opposition à garçon* et *par opposition à fils*) avec une répartition plus homogène, c'est-à-dire sans hiérarchisation nette entre les usages. Toutefois, sa polysémie l'a isolé dans les résultats : *filles* présente un taux de cohésion distributionnelle avec les autres termes de la famille plus faible que celle observée pour la majorité des autres termes. Cela montre l'importance d'exclure le mot *femme* de leurs analyses afin de ne pas biaiser les résultats.

Ainsi, ce mémoire a pour objectif de traiter la polysémie du mot *femme*, partagé entre son sens générique (*être humain de sexe féminin*) et son sens relationnel (*épouse*), que je désignerai désormais par les étiquettes *femme_humain* et *femme_épouse*. L'objectif est alors d'automatiser leur désambiguïsation afin de permettre la réintégration du mot *femme* dans l'analyse distributionnelle du lexique de la parenté. Plus spécifiquement, l'enjeu est de déterminer quels indices linguistiques permettent de lever cette ambiguïté, pour ensuite rendre possible l'élaboration d'un système d'identification et de classement

¹ Représentation numérique d'un élément dans un espace vectoriel.

² Regroupement des mots en thèmes (topics).

³ Demande directe à un LLM de proposer une caractérisation thématique des termes familiaux.

automatique de différentes occurrences issues de corpus. Ce mémoire vise également à observer les usages de *femme_humain* et de *femme_épouse* dans différents corpus, afin d'analyser leur variation selon les genres discursifs et leur évolution dans le temps. Ce second objectif permettra notamment de vérifier l'hypothèse formulée par Tanguy et al. (2024, 2025), selon laquelle le sens *femme_humain* est majoritaire dans les corpus. Il s'agira aussi d'observer si cette répartition varie selon les genres discursifs et si l'usage de *femme_épouse* tend à évoluer par rapport à celui d'*épouse* dans le corpus contemporain de *Frantext*.

Dans cette perspective, la démarche adoptée repose d'abord sur une analyse linguistique des régularités d'emploi du mot *femme*, puis sur leur formalisation dans un système automatique de désambiguïsation. Ce système fait ensuite l'objet d'une évaluation, avant d'être appliqué à plusieurs corpus afin d'observer la répartition des deux sens et leur évolution diachronique.

Ce mémoire est divisé en cinq parties. La première partie est consacrée à l'état de l'art autour de la polysémie, de son traitement automatique et de l'évolution sémantique du mot *femme*. La deuxième partie analyse les régularités d'emploi du mot *femme* à partir des dictionnaires, de *Frantext* et de *Word Sketch Difference*. La troisième partie présente la méthodologie adoptée pour concevoir le système automatique de désambiguïsation, ainsi que la formalisation des règles linguistiques utilisées. La quatrième partie évalue les performances de ce système à partir d'un *gold standard* constitué avec le corpus OPUS *OpenSubtitles*, puis les compare à celles d'un LLM. Enfin, la cinquième partie applique le système à plusieurs corpus afin d'observer la variation des usages de *femme* selon les genres discursifs, ainsi que leur évolution diachronique dans *Frantext*.

Partie 1. La polysémie : définitions et impacts sur le Traitement Automatique des Langues (TAL)

Cet état de l'art s'organise en trois grandes sections. La première examine la notion de polysémie : ses caractéristiques, sa distinction avec l'ambiguïté, l'homonymie et la monosémie, les différents types de relations sémantiques qu'elle implique, ainsi que les débats qu'elle suscite au sein de la communauté scientifique quant à sa définition et à la représentation de ses sens dans les dictionnaires.

La deuxième section s'intéresse au traitement automatique de la polysémie à travers la tâche de désambiguïsation lexicale (*Word Sense Disambiguation, WSD*). Elle présente les principales ressources mobilisées, les approches existantes (symboliques, vectorielles et hybrides), ainsi que les métriques permettant d'évaluer leurs performances.

Enfin, la troisième section se concentre sur le mot *femme*, en retraçant l'évolution de ses définitions dans les dictionnaires et la manière dont les représentations sociales et culturelles qui lui sont associées influencent ses différentes acceptions, aussi bien dans les dictionnaires francophones qu'anglophones.

1.1. La polysémie : concepts-clés

Le terme *polysémie* a été introduit à la fin du XIX^e siècle par Bréal (1897, cité dans Victorri et Fuchs, 1996) pour définir un mot qui *prend un sens nouveau*. Selon Gries (2019, p. 23), « The probably most widely accepted definition of polysemy is as the form of ambiguity where 2+ related senses are associated with the same word ». Cependant, nous le verrons, cette définition ne va pas sans soulever un certain nombre de questions.

1.1.1. Caractéristiques de la polysémie

La polysémie est un phénomène transversal à l'ensemble des catégories grammaticales (verbes, noms, adjectifs, adverbes) et se retrouve dans toutes les langues naturelles (Victorri et Fuchs, 1996). Il convient toutefois de noter qu'un mot polysémique dans une langue ne l'est pas nécessairement dans une autre, ce que nous approfondirons dans la section 1.2.6 à travers les travaux de Hauer et Kondrak (2020), dont l'approche consiste précisément à exploiter la traduction pour désambiguïser les unités polysémiques. Par ailleurs, sur le plan temporel, ces différents sens découlent les uns des autres. En effet, la polysémie d'un mot est un processus diachronique, où le nouveau sens vient en complémentarité de l'ancien (Bréal, 1897).

Ces différents sens ne sont donc pas indépendants les uns des autres. Bertin (2018) souligne qu'ils entretiennent des liens entre eux, ce qui rejoint les travaux de Picoche (1989) autour de la notion de *signifié de puissance*, c'est-à-dire un noyau de sens commun à tous les usages d'un mot. Elle prend l'exemple du mot *créneau*, dont le sens premier désigne une ouverture dans *une muraille de fortification*. À partir de là, deux sens dérivés apparaissent : *faire un créneau* pour se garer entre deux voitures, et *monter au créneau* pour défendre une position. Ces usages ne sont pas des sens séparés et indépendants, ils partagent un même noyau sémantique.

C'est précisément ce noyau partagé qui rend le sens d'un mot polysémique fondamentalement dépendant du contexte. Victorri et Fuchs (1996) décrivent la construction du sens d'un énoncé comme un processus dynamique et réciproque. Le sens global dépend des expressions qui le composent, et ces expressions

sont elles-mêmes influencées par le sens global de l'énoncé dans lequel elles s'inscrivent. Dans cette perspective, un mot polysémique ne dispose pas d'un sens fixe et autonome, c'est son interaction avec les éléments environnants qui oriente l'interprétation vers l'un de ses sens possibles. Comme le formule Fuchs (1996), le problème posé par la polysémie ne concerne pas uniquement le mot lui-même, mais surtout l'interprétation du sens qu'il prend selon le contexte dans lequel il apparaît. Le sens dépend alors des relations établies entre ce mot et les éléments linguistiques qui l'entourent.

1.1.2. Ambiguïté, homonymie et monosémie

Cette dépendance au contexte invite à distinguer la polysémie d'un phénomène qui lui est souvent associé : l'ambiguïté. Pour Jacqueline Picoche (1989), il existe une différence entre la polysémie et l'ambiguïté. La polysémie concerne un mot possédant plusieurs sens possibles et relève de la langue, tandis que l'ambiguïté survient lorsque le sens n'est pas clair dans une phrase donnée et relève donc du contexte. Autrement dit, un mot polysémique n'est pas nécessairement ambigu, car la situation d'énonciation, les connaissances partagées et les intentions du locuteur orientent généralement l'interprétation vers un sens précis. Fuchs (1996) affine cette distinction en opposant l'ambiguïté virtuelle, que le contexte suffit le plus souvent à lever, à l'ambiguïté effective, qui persiste malgré un contexte explicite.

La polysémie ne se définit pas seulement par rapport à l'ambiguïté, mais aussi par opposition à deux phénomènes voisins : l'homonymie et la monosémie. La polysémie est un phénomène beaucoup plus répandu que l'homonymie, car la plupart des mots peuvent, à des degrés divers, être considérés comme polysémiques (Haber et Poesio, 2024). En effet, selon Durkin et Manning (1989, cités dans Haber et Poesio, 2024), environ 40 % des mots fréquents en anglais présentent plusieurs sens, tandis que l'homonymie serait uniquement présente dans 5% des mots de la langue (Victorri et Fuchs, 1996).

Selon Picoche (1989), *la polysémie commence où s'arrête l'homonymie* : il y a polysémie dès lors qu'un lien, même minimal, peut être établi entre les différents emplois d'un mot. Victorri et Fuchs (1996) distinguent ainsi trois cas : la polysémie, lorsque les différents sens d'une même forme lexicale entretiennent des liens sémantiques entre eux ; l'homonymie, où les sens sont indépendants et la forme commune serait *accidentelle*, comme pour le mot *avocat*, qui désigne aussi bien l'homme de loi que le fruit (Bertin, 2018) ; et la monosémie, dans laquelle le sens ne se divise pas, à l'image du mot *tournevis* (Victorri et Fuchs, 1996).

Du point de vue du récepteur, Fuchs (1996) affine cette distinction : face à l'homonymie, le récepteur doit identifier quelle forme est en jeu, tandis que face à la polysémie, il doit choisir entre plusieurs significations associées à un même mot. La différence n'est donc pas seulement structurelle, elle engage aussi le traitement interprétatif.

Afin de distinguer la polysémie de l'homonymie, plusieurs critères ont été proposés en linguistique. Parmi eux, Victorri et Fuchs (1996) s'intéressent aux relations paraphrastiques et aux possibilités de substitution d'un mot selon le contexte. Les auteurs soulignent toutefois que le phénomène de synonymie pure demeure *exceptionnel*, puisqu'un remplacement par un autre terme entraîne presque toujours une légère différence de sens. Ainsi, un mot polysémique n'admet pas nécessairement les mêmes substitutions lexicales dans tous les énoncés. Les variations de paraphrases et de relations sémantiques peuvent alors constituer un indice permettant d'identifier différents emplois polysémiques.

1.1.3. Les différents types de polysémie

Toutes les tâches de désambiguïsation ne présentent pas le même niveau de difficulté, celui-ci dépendant notamment du degré de cohésion sémantique entre les différents sens d'un mot. Dans cette perspective, des chercheurs (Apresjan, 1974, Vicente et Falkum, 2017, cités dans Haber et Poesio, 2024) montrent qu'il n'existe pas qu'une seule forme de polysémie. Parmi les processus sémantiques les plus connus, Haber et Poesio (2024) présentent la métaphore, qui est le transfert d'un sens d'un domaine à l'autre, comme *la bouche d'une rivière* ou la métonymie, qui est le changement de sens à l'intérieur même d'un domaine, comme *l'école*, qui peut être un bâtiment, une institution ou une administration. D'autres (Ortega-Andrés et Vicente, 2019 ; Ortega-Andrés, 2021, cités dans Haber et Poesio, 2024) vont plus loin et proposent une structure hiérarchique pour les mots polysémiques. Pour l'exemple du mot *école*, il y a un noyau abstrait, qui est l'institution, des réalisations sociales, comme l'organisation, le personnel ou les étudiants, ainsi que des réalisations temporelles et physiques, telles que l'emploi du temps ou le bâtiment.

La métaphore occupe une place centrale dans l'étude de la polysémie, dans la mesure où elle participe à la création de nouveaux sens (Lakoff et Johnson, 1980, cités dans Bertin, 2018). Elle a été très utilisée par les sémanticiens pour décrire la polysémie de nombreux noms, dont les emplois les plus concrets sont généralement considérés comme premiers, tandis que les emplois plus abstraits dérivent progressivement de ces usages initiaux (Pauly, 2010, cité dans Bertin, 2018).

1.1.4. Les limites et les débats autour de cette notion

Cette importance accordée aux mécanismes de création et d'organisation des sens montre toutefois que la polysémie reste une notion difficile à définir de manière claire et stable. Malgré les nombreux travaux consacrés à ce phénomène, les chercheurs ne s'accordent pas toujours sur la manière de définir les différents sens d'un mot ni sur les relations qu'ils entretiennent entre eux (Mazaleyrat, 2010).

Par exemple, Bertin (2018) explique que si on se base uniquement sur l'intuition pour dire qu'un mot est polysémique, on perd en rigueur scientifique. D'ailleurs, certains chercheurs (Enfield, 2002, cité dans Bertin, 2018) remettent en question l'usage même de ce terme, car ils considèrent que l'on invente des relations sémantiques *a posteriori* pour justifier le caractère polysémique d'un terme. De même, selon Goddard (2000, cité dans Bertin, 2018), le fait d'expliquer tous les usages d'un mot par un seul sens commun revient justement à de la monosémie. Par exemple, le fait qu'il existe plusieurs formes d'amour ne permettrait pas de postuler que le mot *love* soit polysémique.

Ce désaccord se manifeste également dans la manière de dénombrer et de regrouper les sens. En effet, Gries (2019) distingue deux approches dans l'étude de la polysémie. D'un côté, les *lumpers*, qui estiment qu'un mot possède un petit nombre de définitions distinctes et beaucoup d'usages différents appartenant au même sens général. De l'autre, les *splitters*, qui considèrent au contraire qu'un mot a de nombreux sens distincts même s'ils sont proches.

Toutefois, dans le cadre de ce mémoire, nous nous en tiendrons à la définition de Gries (2019), citée en début de cette section.

1.1.5. La représentation des sens dans les dictionnaires

La distinction entre polysémie et homonymie dans les dictionnaires ne fait pas non plus consensus (Bertin, 2018). La représentation des différents sens d'un mot dépend en effet des choix théoriques et méthodologiques adoptés par les lexicographes. Bertin (2018) souligne notamment qu'un mot polysémique est plus difficile à définir qu'un mot monosémique, ce qui conduit généralement à des articles lexicographiques plus longs et plus complexes. De plus, les dictionnaires proposent souvent une liste d'acceptions distinctes qui ne reflète pas toujours le fonctionnement réel du sens en contexte, celui-ci étant plus progressif et fortement dépendant de l'usage.

Ces difficultés de représentation influencent directement la manière dont les dictionnaires organisent les différents sens d'un mot. Ainsi, Victorri et Fuchs (1996) expliquent que les dictionnaires consacrent généralement plusieurs articles aux mots homonymiques, un seul article subdivisé aux mots polysémiques, et un article unique sans subdivision aux mots monosémiques.

La question du sens premier est également sujette à débat, plusieurs critères pouvant être mobilisés pour hiérarchiser les acceptions d'un mot polysémique : l'intuition du lexicographe, le degré de concrétude du sens, l'origine étymologique, la fréquence d'usage en corpus, ou encore le sens prototypique (Mazaleyrat, 2010). Gold (1986) explique ainsi que certains lexicographes choisissent d'organiser les sens selon leur importance supposée, en plaçant les acceptions les plus fréquentes en début d'article afin de faciliter la consultation. Pour Bertin (2018), les sens les plus concrets apparaissent généralement en premier, ceux-ci correspondant souvent aux usages les plus fréquents avant les emplois dérivés ou plus abstraits.

Cependant, cette structuration a posé certaines limites dans une perspective de traitement automatique des langues, où la pluralité des sens doit être formalisée de manière explicite afin de pouvoir être exploitée par des systèmes automatiques (Mathieu-Colas, 1995).

1.1.6. Traitement cognitif des sens multiples

Face à cette omniprésence de sens multiples, il est nécessaire de comprendre comment ils s'organisent sur le plan conceptuel. Gries (2019) utilise le terme de *catégories radiales*, notion initialement proposée par Lakoff (1987), lorsqu'un mot a un sens central ainsi que plusieurs extensions liées par des similarités. L'auteur prend l'exemple du mot *mère*, qui renvoie à *une femme qui a donné naissance à un enfant*. *Mère porteuse*, *mère adoptive* ou *mère patrie* ont un sens qui gravite également autour de la maternité, tout en s'en éloignant un peu.

Au-delà de la description de ces structures, il est intéressant d'observer comment elles sont traitées sur le plan cognitif par le locuteur. En psycholinguistique, la polysémie n'était pas forcément étudiée dès le départ, on parlait plutôt d'ambiguïté, mais cela ne distinguait pas la polysémie de l'homonymie. Cependant, les études ont montré que les sens d'un mot polysémique partagent une représentation mentale partielle commune, alors que les homonymes sont traités comme des mots différents (Gries, 2019 ; Haber et Poesio, 2024). Cet aspect est important à prendre en compte pour comprendre le fonctionnement cognitif de la polysémie.

Finalement, ces recherches illustrent la complexité à définir la polysémie ainsi que la diversité des relations sémantiques qu'elle englobe. En effet, les sens entretiennent entre eux des liens hiérarchiques, métaphoriques ou métonymiques, qui varient selon les contextes d'usage et les cadres cognitifs

mobilisés. Cette variabilité rend difficile toute définition unique et stable de la polysémie, tant du point de vue linguistique que computationnel. C'est précisément cette complexité qui nous amène à examiner comment le TAL appréhende ce phénomène.

1.2. Le traitement automatique de la polysémie

En TAL, la différence entre polysémie et homonymie est moins centrale qu'elle ne l'a été en linguistique ou en psychologie. Cette distinction existe bien en TAL, mais elle a historiquement été ignorée ou simplifiée. En effet, les systèmes traitent souvent toutes les ambiguïtés de la même manière, sans se soucier de leur relation. À titre d'exemple, *WordNet* représente l'ensemble des sens d'un mot de manière uniforme, indépendamment du fait qu'ils relèvent de la polysémie ou de l'homonymie (Haber et Poesio, 2024).

Toutefois, selon Weaver (1949), qui a été le premier à reconnaître le problème de la polysémie dans la traduction automatique, comprendre le sens d'un mot passe par la prise en compte du contexte. Il compare la compréhension d'un texte à l'examen d'un livre à travers une fente : on ne peut saisir le sens d'un mot qu'en voyant également ses voisins. Cette métaphore introduit le principe de désambiguïstation lexicale (*Word Sense Disambiguation*), c'est-à-dire « rendre explicite la sémantique d'un mot dans son contexte en identifiant le sens le plus approprié à partir d'un inventaire de sens prédéfini. » (Bevilacqua et al., 2021, p. 4330). Ce problème est central en traitement automatique des langues, car la compréhension d'un texte, même par une machine, dépend fortement de la capacité à interpréter correctement les mots ambigus.

Navigli (2009) décrit le principe fondamental de tout système traitant cette tâche comme le processus par lequel un système informatique identifie, pour chaque mot potentiellement ambigu, le sens qui s'accorde le mieux avec le contexte dans lequel il apparaît. Au lieu de traiter les mots de manière isolée, le système considère une unité de contexte (par exemple une phrase ou un segment de texte) et évalue les différentes interprétations possibles des mots qui la constituent. Pour effectuer ce choix, les méthodes de WSD exploitent diverses formes de connaissance linguistique ou sémantique : corpus de textes, qu'ils soient bruts ou annotés en sens, lexiques électroniques, dictionnaires ou réseaux sémantiques.

1.2.1. Les ressources pour la désambiguïstation

Parmi les sources de connaissances utilisées pour traiter la polysémie, il existe des inventaires de sens, c'est-à-dire des ensembles qui listent les différents sens possibles d'un mot. La tâche de désambiguïstation s'appuie aussi sur des corpus annotés, notamment pour entraîner et tester les programmes (Bevilacqua et al., 2021). Parmi les jeux de données les plus utilisés, on trouve le corpus *SemCor* pour l'entraînement (Miller et al., 1993, cités dans Bevilacqua et al., 2021), ainsi que les jeux de données issus des campagnes d'évaluation internationales *Senseval* et *SemEval* pour le test. Ces données sont surtout en anglais, mais il existe un *benchmark* plus récent, le *XL-WSD*, qui a introduit des données pour 18 autres langues (Pasini et al., 2021, cités par Bevilacqua et al., 2021).

Le corpus *SemCor* est l'une des ressources les plus importantes pour l'entraînement des systèmes de désambiguïstation lexicale, il a été développé dans le cadre des travaux de Miller et al. (1993). Dans *SemCor*, l'annotation est réalisée manuellement à partir de textes du *Brown Corpus*, les sens retenus étant ceux de *WordNet*. Les statistiques associées à *SemCor* varient selon les versions et les critères de comptage, mais la version *SemCor 3.0*, aujourd'hui la plus couramment utilisée, est constituée de 352

textes issus du *Brown Corpus*. Toutefois, *SemCor* reflète les choix de granularité de *WordNet*, souvent critiqués pour être trop fins (Kilgarriff, 1997), comme nous le verrons ensuite.

En ce qui concerne les inventaires de sens, les plus utilisés sont *WordNet* et *BabelNet*. Le premier est une base de données lexicographique de l'anglais, organisée en ensembles de synonymes (*synsets*), où chaque synonyme correspond à un sens. Pour chaque *synset*, *WordNet* fournit des définitions et des exemples, et ces *synsets* sont reliés entre eux par des relations lexicales, principalement l'hyponymie. Le second, *BabelNet*, est un réseau sémantique multilingue combinant plusieurs ressources lexicographiques et encyclopédiques, organisé semblablement à *WordNet*. La Figure 1 est la tentative de Caldarola et al. (2016) de représenter *WordNet* sous forme visuellement lisible pour un humain à l'aide du logiciel Cytoscape. Elle illustre les différents *synsets* associés au mot *home*, où chaque nœud correspond à un sens distinct relié à d'autres unités lexicales, mettant ainsi en évidence l'organisation en réseau des sens dans *WordNet*.

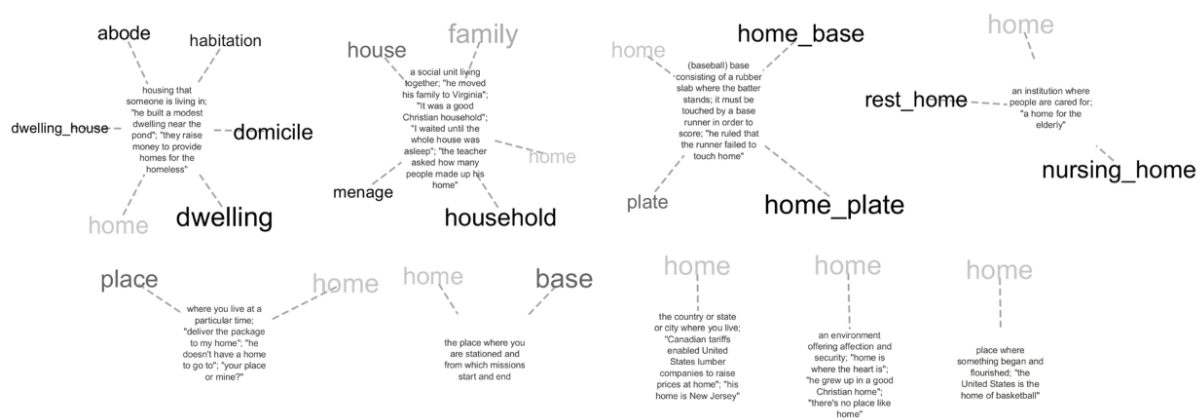


Figure 1 : Synsets *WordNet* contenant le mot « home », avec un style personnalisé dans *Cytoscape* (Caldarola et al. (2016)).

Cependant, les inventaires de sens ont un problème de granularité trop fine, en d'autres termes, ils séparent les sens de manière trop détaillée. En effet, les différentes significations d'un même lexème sont difficiles à distinguer et chaque interprétation d'un mot est simplement stockée comme un sens différent sans tenir compte du lien entre eux (Haber et Poesio, 2024). De plus, les inventaires de sens énumèrent la signification dans une liste finie, bien que la langue soit en constante évolution (Bevilacqua et al., 2021). Cependant, ces limites ne relèvent pas uniquement de choix techniques ou de modélisation, mais interrogent plus profondément la manière dont le sens des mots est conceptualisé en traitement automatique des langues.

1.2.2. La notion de sens en TAL

La tâche de désambiguïsation lexicale repose sur une hypothèse centrale : celle que les mots possèdent un ensemble fini de sens distincts qu'il est possible d'identifier et de catégoriser. Cependant, cette hypothèse est loin de faire consensus. Kilgarriff (1997) remet en question leur existence en affirmant qu'il ne croit pas aux sens des mots. Selon lui, les distinctions proposées dans les dictionnaires ou les ressources comme *WordNet* ne reflètent pas nécessairement la réalité du langage, mais résultent plutôt de choix arbitraires faits par les lexicographes.

Dans cette perspective, le sens d'un mot ne correspond pas à un ensemble fini d'entrées distinctes, mais à un continuum de variations dépendant du contexte d'usage. Cela explique en partie pourquoi les

annotateurs humains ont souvent des difficultés à s'accorder sur le choix d'un sens précis, et pourquoi la tâche de désambiguïsation lexicale elle-même peut être vue comme artificielle. Cette remise en question a conduit certains travaux récents à privilégier des approches contextuelles continues, comme les *embeddings* contextualisés, qui ne reposent plus sur des inventaires de sens prédéfinis.

Ainsi, la critique de Kilgarriff (1997) remet en cause le fondement même de la WSD classique, qui repose sur l'hypothèse que les sens sont des catégories discrètes identifiables. Comme je le montrerai dans la section 2.1 à travers l'analyse du mot *femme*, les distinctions proposées par les dictionnaires varient selon les ressources et ne permettent pas toujours de rendre compte des usages en contexte. Certains exemples, tels que *femme veuve* ou *bonne femme*, illustrent la difficulté à attribuer un sens unique et stable, confirmant ainsi le caractère non discret et contextuel du sens lexical.

1.2.3. Approche symbolique : l'algorithme de Lesk

Malgré ces critiques, de nombreuses approches de désambiguïsation lexicale reposaient sur l'hypothèse de sens discrets issus d'inventaires comme *WordNet*. Ces approches, dites symboliques, exploitaient explicitement les définitions lexicales pour déterminer le sens d'un mot en contexte. L'un des algorithmes les plus représentatifs de cette approche est l'algorithme de Lesk (1986). Dans les versions simplifiées, le principe consistait à comparer directement le contexte du mot cible avec les différentes définitions proposées dans un dictionnaire, en ne retenant que les mots porteurs de sens. L'ordre des mots n'était pas pris en compte : il s'agissait simplement de mesurer le chevauchement lexical entre le contexte et chaque définition. Le sens retenu était alors celui dont la définition présente le plus grand chevauchement lexical avec les mots du contexte (Navigli, 2009).

Par exemple, dans la phrase « I deposited my money at the *bank* », le mot *bank* peut désigner un établissement financier ou le bord d'une rivière. L'algorithme sélectionnera le sens dont la définition présente le plus de chevauchements lexicaux avec le contexte, ici le sens *établissement financier*, c'est-à-dire une institution qui accepte les dépôts d'argent et les utilise notamment pour des activités de prêt, en raison de la présence du mot *money*.

Cette approche comportait certaines limites. Par exemple, plus la phrase était longue, plus il y avait de définitions à croiser entre elles, ce qui était compliqué à l'époque à cause de la puissance des ordinateurs. Des variantes ont été proposées pour améliorer cette méthode, notamment l'algorithme de Lesk étendu (Banerjee et Pedersen, 2003, cités dans Raganato et al., 2017), qui enrichit les définitions en intégrant des relations issues de *WordNet*. Toutefois, ces approches restent limitées, car elles reposent sur un appariement de surface et ne capturent pas la similarité sémantique entre les mots.

Bien que leur efficacité soit aujourd'hui dépassée par les approches neuronales, les méthodes symboliques restent importantes dans l'histoire de la désambiguïsation lexicale. Leurs limites tiennent notamment au fait qu'elles reposent fortement sur des inventaires de sens prédéfinis et sur une représentation discrète du sens, dont les difficultés ont été mises en évidence précédemment. Ces limites ont conduit au développement de nouvelles méthodes visant à représenter le sens de manière plus flexible et graduelle.

1.2.4. Représentations vectorielles des mots

Dans cette perspective, plusieurs approches ont cherché à représenter les mots à partir de leurs contextes d'apparition plutôt qu'à partir de définitions lexicales. En TAL, cela a conduit au développement de représentations vectorielles des mots. Apidianaki (2023) distingue ainsi trois grandes méthodologies, qui ont émergé progressivement avec les avancées technologiques, pour représenter les mots sous forme de vecteurs dans un espace sémantique.

Dans cette approche, chaque mot est converti en un vecteur numérique, plus précisément une position dans un espace à plusieurs dimensions. Mikolov et al. (2013) expliquent que, dans cet espace de représentations vectorielles, les mots ayant des sens proches se trouvent à proximité les uns des autres : la distance entre deux vecteurs reflète leur similarité sémantique. Ainsi, plus deux mots sont liés par le sens, plus leurs vecteurs sont proches. Par exemple, ils démontrent qu'en supprimant le trait *homme* à *roi*, et en ajoutant le trait *femme*, on obtient un vecteur très proche de celui de *reine*.

Dans la première méthode présentée par Apidianaki (2023), appelée *Static Word Embeddings*, le modèle apprend de manière non supervisée à partir de corpus de textes et attribue à chaque mot un vecteur unique, identique dans tous les contextes. Cependant, le défaut est que si un mot est polysémique, le modèle mélange tous ses sens dans une seule représentation.

Ensuite, avec la méthode *Contextualized Embeddings*, il est possible d'avoir un vecteur différent pour chaque occurrence du mot en fonction du contexte. En effet, ces modèles génèrent dynamiquement un vecteur en fonction du contexte précis de la phrase, ce qui réduit naturellement le problème de la polysémie. C'est comme cela que fonctionnent certains modèles comme ELMo, BERT ou mBERT.

Enfin, la méthode *Sense-aware Embeddings* a permis de calculer plusieurs vecteurs par mot, un pour chaque sens possible. C'est cette méthode qui nous intéresse pour traiter la polysémie, car ces modèles ont justement été conçus pour la gérer (Apidianaki, 2023). Cette approche générale se décline en trois sous-méthodes distinctes selon la manière d'identifier les sens. Premièrement, la méthode des *Multi-prototype Embeddings* consiste à apprendre plusieurs vecteurs pour un même mot, chacun correspondant à un usage particulier. Ces vecteurs sont obtenus en regroupant (par *clustering*) les occurrences du mot ayant des contextes similaires, ce qui permet d'identifier automatiquement différents sens. Deuxièmement, les méthodes basées sur la traduction utilisent les traductions dans d'autres langues comme indicateurs de sens. Un mot polysémique en langue source peut avoir plusieurs traductions selon son sens. Chaque traduction correspond à un sens différent du mot original, et si plusieurs traductions sont synonymes entre elles, elles décrivent le même sens. Troisièmement, la méthode *Sense Embeddings* produit des vecteurs pour les sens trouvés dans les ressources lexicographiques, en combinant parfois ces connaissances avec des informations provenant de corpus de textes.

Ces différentes méthodes permettent ainsi de représenter les différents sens d'un mot sous forme vectorielle. Ces représentations constituent alors la base sur laquelle s'appuient les modèles de désambiguïsation lexicale, dont l'objectif est de sélectionner le sens le plus approprié dans un contexte donné.

1.2.5. Les approches principales de la WSD

Une fois les mots représentés sous forme de vecteurs, il reste à déterminer le sens précis dans un contexte donné. C'est l'objectif de la désambiguïsation lexicale (WSD), qui met en œuvre trois approches

principales pour construire des modèles, qui exploitent ces représentations. Tout d'abord, il y a les modèles supervisés, qui sont les plus efficaces. Ils traitent la tâche comme un problème de classification en entraînant des réseaux de neurones (souvent basés sur des modèles comme BERT) à associer le vecteur contextuel d'un mot à son étiquette de sens, en apprenant à partir de corpus annotés manuellement comme *SemCor*.

Ensuite, les approches basées sur la connaissance exploitent les inventaires de sens comme *WordNet*. Elles utilisaient historiquement des algorithmes de graphe (comme *PageRank*) pour trouver quel sens est le plus connecté par rapport aux mots du contexte. Toutefois, les méthodes récentes comparent désormais directement le vecteur du mot avec les vecteurs des définitions (gloses) ou des exemples fournis par le dictionnaire.

Enfin, les approches hybrides combinent les deux stratégies précédentes pour obtenir les meilleurs résultats. Elles s'entraînent sur des exemples comme les modèles supervisés et s'aident des informations du graphe de connaissances, comme les relations sémantiques ou les définitions, ce qui permet de mieux identifier les sens rares (Bevilacqua et al., 2021).

1.2.6. Approches basées sur la traduction

Nous avons vu dans la section 1.1.1 que l'un des moyens de désambiguïser consiste à passer par la traduction. Des méthodes récentes exploitent justement les données multilingues, comme l'hypothèse de Hauer et Kondrak (2020, cités dans Haber et Poesio, 2024). Cette hypothèse permet de classer les sens dans des ressources comme *BabelNet* : si deux sens d'un mot se traduisent par le même terme dans au moins une autre langue, ils sont considérés comme polysémiques, sinon comme homonymiques. Ces approches sont particulièrement intéressantes pour traiter la granularité parfois trop fine des inventaires de sens.

Leur méthodologie est de transformer les traductions multilingues fournies par *BabelNet* en vecteurs de caractéristiques, où chaque dimension d'un vecteur correspond à une langue ou une traduction possible. Puis, ils comparent ces vecteurs avec une mesure de similarité pour décider si deux sens sont proches (polysémie) ou éloignés (homonymie). Par exemple, pour le mot *femme*, si l'anglais sépare les sens (*woman / wife*), l'espagnol les réunit (*mujer / mujer*). La présence d'une traduction commune dans au moins une langue permet ainsi de classer correctement ces sens comme polysémiques, offrant une manière plus objective de caractériser l'ambiguïté lexicale.

Au-delà de cette analyse théorique, les auteurs proposent également une approche concrète pour améliorer les systèmes de désambiguïstation lexicale en intégrant directement l'information issue de la traduction automatique (Luan et al., 2020). Leur méthode combine un système de traduction neuronale avec des ressources comme *BabelNet* afin de relier les traductions aux sens possibles d'un mot. Le principe repose sur l'idée que différentes traductions d'un mot reflètent souvent des sens différents : en identifiant la traduction du mot cible dans une ou plusieurs langues, il est possible de restreindre ou d'orienter le choix du sens. Deux stratégies sont alors mises en œuvre : une approche dite *hard constraint*, qui impose des contraintes strictes basées sur les traductions, et une approche *soft constraint*, qui les intègre de manière probabiliste. Cette dernière se montre plus flexible, car elle considère les traductions comme des indices plutôt que comme des correspondances exactes.

Les résultats montrent que l'intégration de l'information multilingue améliore significativement les performances des systèmes de WSD. Par exemple, l'approche proposée permet d'atteindre une F-mesure

allant jusqu'à 80.1 sur le jeu de test SemCor, notamment lorsque plusieurs langues (comme l'italien et le japonais) sont combinées. Ces performances constituent, au moment de la publication, les meilleurs résultats obtenus en exploitant la traduction automatique pour la désambiguïsation lexicale. Cette approche met ainsi en évidence le rôle central des traductions comme source d'information sémantique pour la WSD, en montrant qu'elles peuvent enrichir efficacement les systèmes existants.

1.2.7. Évaluation des performances

Pour évaluer les performances en traitement automatique des langues, Habert (2005) et Navigli (2009) présentent plusieurs métriques issues de la littérature, telles que la précision, le rappel et la F-mesure. Bien que ces métriques soient générales et applicables à de nombreuses tâches du TAL, elles sont également utilisées dans le cadre de la désambiguïsation lexicale, et elles serviront également à l'évaluation dans la quatrième partie de ce mémoire.

Le rappel mesure la capacité du système à identifier l'ensemble des sens corrects : l'absence de certaines bonnes réponses correspond à du silence. La précision, quant à elle, calcule la proportion de réponses correctes parmi celles produites par le système : la présence de réponses erronées correspond alors à du bruit. Enfin, le calcul de la F-mesure représente la moyenne harmonique entre la précision et le rappel (Navigli, 2009), et constitue aujourd'hui le standard de référence dans les campagnes d'évaluation internationales.

Ces performances peuvent ensuite être interprétées par rapport à un minimum, la *random baseline*, et un maximum, l'accord inter annotateurs. La *random baseline* correspond à la performance obtenue si l'on choisit un sens au hasard parmi ceux disponibles pour chaque mot (Navigli, 2009). Par exemple, pour le mot *femme*, qui a deux sens principaux (*être humain de sexe féminin* et *épouse*), la *random baseline* serait de 50 % puisqu'il y aurait une chance sur deux d'obtenir le bon sens. Elle sert uniquement de point de référence minimal, tandis que l'accord inter annotateurs est un maximum réaliste réalisé par des humains.

Comme le montre Navigli (2009), la comparaison des systèmes de WSD a longtemps été compliquée par la diversité des jeux de tests, des inventaires de sens et des ressources utilisées. Cette hétérogénéité rendait les résultats difficilement comparables, en particulier avant les premières campagnes de standardisation comme *Senseval*. Pour répondre à ce problème, Raganato et al. (2017) ont proposé un cadre d'évaluation unifié (*Unified Evaluation Framework*). Ce cadre permet aujourd'hui de comparer plus facilement des approches classiques, comme l'algorithme de Lesk, avec des modèles plus récents, dans un environnement d'évaluation commun et cohérent. Ce cadre permet notamment de produire des comparaisons fiables entre systèmes, comme l'illustre le Tableau 1.

Ce tableau présente les performances en F-mesure de différents systèmes de désambiguïsation lexicale sur plusieurs jeux de données standards (*Senseval-2*, *Senseval-3*, *SemEval-07*, *SemEval-13* et *SemEval-15*). Les systèmes sont regroupés en deux grandes catégories : les approches supervisées, entraînées sur des corpus annotés (*SemCor* ou *SemCor* + *OMSTI*), et les approches fondées sur la connaissance (*knowledge-based*), qui utilisent uniquement des ressources lexicales comme *WordNet*.

	Tr. Corpus	System	Senseval-2	Senseval-3	SemEval-07	SemEval-13	SemEval-15
Supervised	SemCor	IMS	70.9	69.3	61.3	65.3	69.5
		IMS+emb	71.0	69.3	60.9	67.3	71.3
		IMS _s +emb	72.2	70.4	62.6	65.9	71.5
		Context2Vec	71.8	69.1	61.3	65.6	71.9
		MFS	65.6	66.0	54.5	63.8	67.1
		<i>Ceiling</i>	<i>91.0</i>	<i>94.5</i>	<i>93.8</i>	<i>88.6</i>	<i>90.4</i>
	SemCor + OMSTI	IMS	72.8	69.2	60.0	65.0	69.3
		IMS+emb	70.8	68.9	58.5	66.3	69.7
		IMS _s +emb	73.3	69.6	61.1	66.7	70.4
		Context2Vec	72.3	68.2	61.5	67.2	71.7
		MFS	66.5	60.4	52.3	62.6	64.2
		<i>Ceiling</i>	<i>91.5</i>	<i>94.9</i>	<i>94.7</i>	<i>89.6</i>	<i>91.1</i>
Knowledge	-	Lesk _{ext}	50.6	44.5	32.0	53.6	51.0
		Lesk _{ext} +emb	63.0	63.7	56.7	66.2	64.6
		UKB	56.0	51.7	39.0	53.6	55.2
		UKB _{gloss}	60.6	54.1	42.0	59.0	61.2
		Babelfy	67.0	63.5	51.6	66.4	70.3
		WN 1 st sense	66.8	66.2	55.2	63.0	67.8

Tableau 1 : F-mesure (%) pour différents modèles dans cinq ensembles de données WSD contenant tous les mots (Raganato et al., 2017).

On observe d’abord que les systèmes supervisés obtiennent globalement de meilleurs résultats que les approches *knowledge-based*. Par exemple, les modèles comme *IMS* ou *Context2Vec* dépassent régulièrement les 65-70 de F-mesure, alors que les méthodes comme *Lesk* ou *UKB* restent nettement en dessous. Ce constat confirme que les approches apprenant à partir de données annotées sont plus efficaces pour la tâche de WSD. Le tableau montre aussi que certaines améliorations, comme l’ajout d’*embeddings* ou l’utilisation de modèles neuronaux, permettent de gagner légèrement en performance, sans pour autant provoquer de rupture majeure. Enfin, on remarque que la baseline du sens le plus fréquent (*Most Frequent Sense (MFS)*) reste difficile à dépasser, en particulier pour les systèmes fondés sur la connaissance. Cela souligne l’importance de la fréquence des sens dans les données et les limites des approches reposant uniquement sur des ressources lexicales.

Dans l’article de Raganato et al. (2017), les systèmes de désambiguïsation lexicale évalués reposent principalement sur trois grandes approches : des modèles supervisés classiques, des modèles avec *embeddings* statiques, ainsi que des approches *knowledge-based*. Ces travaux s’inscrivent dans un état de l’art antérieur à l’émergence des modèles de type *Transformer*.

Cependant, même avec l’arrivée de ces modèles plus récents, la WSD ne peut toujours pas être considérée comme un problème résolu. En effet, comme le montrent Loureiro et al. (2021), plusieurs difficultés persistent. D’une part, la création de données annotées de qualité reste un processus long, coûteux et difficile à mettre en œuvre à grande échelle, ce qui limite la couverture des mots et de leurs différents sens. D’autre part, certaines catégories grammaticales, en particulier les verbes, demeurent difficiles à traiter : ils présentent souvent un grand nombre de sens proches, ce qui complique leur désambiguïsation. Ainsi, malgré les progrès apportés par les modèles neuronaux récents, la désambiguïsation lexicale reste un domaine de recherche actif, avec des défis qui restent à relever (Apidianaki, 2023).

1.2.8. La place de la désambiguïsation lexicale dans le contexte des LLM

Les résultats présentés précédemment montrent que les performances en WSD se sont progressivement améliorées avec les approches supervisées, puis avec les modèles neuronaux. Cependant, l’arrivée des grands modèles de langue marque une nouvelle étape dans cette évolution. Contrairement aux systèmes

plus anciens, qui reposaient sur une tâche de désambiguïsation explicitement définie, les LLM peuvent résoudre une partie des ambiguïtés lexicales de manière implicite, à partir du contexte fourni dans la phrase ou dans le texte.

Navigli (2026) revient ainsi sur l'idée selon laquelle les LLM rendraient la WSD obsolète. Il souligne que les LLM les plus récents atteignent désormais des performances comparables à celles de systèmes spécialisés de WSD et peuvent même les dépasser, notamment lorsque la tâche est formulée par prompt, ou lorsque le modèle est adapté par *fine-tuning*⁴.

Cependant, il montre que cette amélioration des performances ne signifie pas que la WSD soit devenue inutile. Au contraire, les travaux récents prouvent que les LLM restent sensibles à plusieurs limites : ils privilégient souvent les sens les plus fréquents, rencontrent davantage de difficultés avec les sens rares et peuvent produire des résultats variables selon la formulation du prompt ou selon la langue considérée. Les erreurs observées en traduction automatique montrent également que la résolution implicite des ambiguïtés lexicales ne suffit pas toujours à garantir une interprétation correcte du sens en contexte.

Dans cette perspective, la désambiguïsation lexicale change donc en partie de rôle avec les LLM. Elle ne sert plus seulement à développer des systèmes capables d'attribuer explicitement un sens à un mot ambigu. Elle permet aussi d'évaluer les modèles eux-mêmes, en observant leur capacité à comprendre le sens d'un mot selon son contexte. Les tâches de WSD restent ainsi utiles pour repérer les limites des LLM, notamment lorsqu'ils sont confrontés à des sens rares, à des contextes ambigus ou à des formulations différentes. La WSD garde donc une place importante dans le contexte actuel, mais son intérêt se déplace : elle devient aussi un moyen d'analyser la compréhension lexicale des systèmes.

Finalement, la désambiguïsation lexicale ne peut pas être considérée comme une tâche obsolète avec l'arrivée des LLM. Même si ces modèles atteignent aujourd'hui des performances élevées et peuvent rivaliser avec des systèmes spécialisés, ils ne parviennent pas à résoudre toutes les ambiguïtés lexicales (Navigli, 2026). De plus, les performances humaines elles-mêmes ne correspondent pas à une réussite totale, puisqu'un annotateur humain atteint 91,3 de F-mesure dans l'évaluation présentée par l'auteur. Cela montre que la WSD reste une tâche complexe, dont l'intérêt est toujours actuel pour évaluer la compréhension du sens en contexte.

1.3. Le mot *femme* : évolution sémantique

Dans la section 2.1 nous verrons de quelle manière les dictionnaires actuels traitent le mot *femme* ainsi que sa polysémie. Avant cela, il est essentiel de comprendre quelles recherches ont déjà été effectuées sur ce sujet et comment ce mot est défini, notamment pour comprendre son évolution.

Benhamou (1986) a abordé la différence de traitement du mot *femme* comparé au mot *homme* dans les dictionnaires. Selon elle, « L'identité féminine sans référence à l'autre n'existe quasiment pas » (p. 44). En effet, elle explique, au travers d'exemples, que dans les diverses définitions de la *femme*, *l'homme* est toujours présent, que ce soit de manière *formelle* (explicite) ou *effacée* (implicite). Par exemple, dans le sens *épouse*, la présence de l'homme est explicite, puisque la femme est l'épouse de l'homme. D'un autre côté, quand la femme est décrite comme un être *délicat* ou *dévoué*, elle l'est de manière plus implicite par rapport à une tierce personne, c'est-à-dire *l'homme* ou *la famille*. Cela permet de comprendre, si on s'intéresse aux entrées du dictionnaire, que *femme* est souvent placée en position

⁴ Technique consistant à spécialiser un modèle d'intelligence artificielle pré-entraîné à l'accomplissement d'une tâche spécifique.

subordonnée à l'*homme*, ce dernier apparaissant comme la référence première. Ainsi, même si on pourrait penser que *femme* est un co-hyponyme d'*homme*, étant tous les deux des hyponymes d'*être humain*, il semble plutôt être un hyponyme du mot *homme* (Baider et Jacquey, 2008).

Cette asymétrie est également relevée par Le Draoulec et Rebeyrolle (2025), qui rapprochent le cas de *femme* de celui de *fille*. Les autrices montrent que, du côté féminin, une même forme lexicale peut renvoyer à l'identité de l'individu mais aussi à une relation familiale ou conjugale : *fille* peut désigner à la fois l'*individu féminin* et la fille de quelqu'un, tandis que *femme* peut désigner à la fois l'*être humain de sexe féminin* et l'*épouse*. À l'inverse, du côté masculin, ces deux dimensions sont distinguées par des termes différents, comme *garçon/fils* ou *homme/mari*. Les autrices expliquent ainsi que la polysémie des mots *fille* et *femme* sont le résultat d'une évolution historique complexe, liée aux représentations sociales dans lesquelles ces emplois se sont développés.

Baider et Jacquey (2008) ont utilisé *WordNet* et ont remarqué que, de manière générale, dans les définitions, la femme (*woman*) est très souvent associée au mariage et ne joue pas de rôle significatif dans la vie de l'homme. Elle a souvent un rôle dévalorisant et réduit aux stéréotypes culturels et traditionnels (Benhamou, 1986 ; Baider et Jacquey, 2008). De plus, Benhamou (1986) nous fait remarquer que les entrées du dictionnaire du terme *femme* sont très souvent dépréciatives, contrairement à *homme*. Cela rejoint les recherches de Baider et Jacquey (2008) qui expliquent que dans les paires minimales, comme *homme facile / femme facile*, on observe des dissymétries interprétatives, où la sexualisation du terme concerne principalement les femmes.

Nous pouvons tout de même constater que les définitions du mot *femme* ont évolué. Par exemple, dans une émission diffusée sur France Inter, la journaliste Christine Siméone récapitule les différentes avancées de la définition de *femme* dans le dictionnaire de l'*Académie française*, et rappelle que dans sa première version, le dictionnaire proposait la définition de *femme* comme étant simplement *la femelle de l'homme*. Quant à l'historienne Muriel Salle (2023), elle démontre que les étiquettes *femme*, *mère* ou *parent qui accouche* ne sont pas neutres et font l'objet de renégociation politique, notamment à cause des questions sur la transidentité. Avec l'exemple d'une campagne sur le Planning familial, elle montre que « les spécificités biologiques définissant habituellement *femmes* et *hommes* » sont aujourd'hui remises en question. Cependant, l'évolution est importante dans les langues et c'est justement une liberté que la polysémie offre aux locuteurs (Victorri et Fuchs, 1996).

Dans ce mémoire, nous cherchons à identifier le sens *épouse* de *femme* et ce sens est également impacté par les changements de notre société. Pour mesurer l'impact des changements sociétaux sur ce concept, il est pertinent d'élargir notre observation à l'anglais. En effet, contrairement au français où le seul terme *femme* englobe tout, l'anglais opère une distinction lexicale entre l'*être humain de sexe féminin* (*woman*) et l'*épouse* (*wife*). Pourtant, malgré l'existence de ces deux termes distincts, les définitions de *woman* et de *wife* ont dû évoluer récemment, prouvant que les représentations sociales influencent le lexique au-delà du simple découpage lexical entre deux sens.

Cela s'observe notamment dans la mise à jour effectuée en 2020 par *Oxford University Press* pour son dictionnaire d'usage courant, l'*Oxford Dictionary of English*⁵. Il s'agit d'un des dictionnaires dont les données sont proposées par les moteurs de recherche comme *Google*. C'est pour cela qu'une polémique

⁵ L'article du *Guardian* (cité par Forgar, S., 2020) renvoie vers l'*Oxford English Dictionary* (dictionnaire historique, <https://www.oed.com/?tl=true>), qui a également été mis à jour. Cependant, la polémique concernant les résultats *Google* visait spécifiquement les données de l'*Oxford Dictionary of English* (dictionnaire d'usage courant), fournies par *Oxford Languages*.

a éclaté : les gens avaient accès à ces exemples directement avec une recherche sur internet. Ainsi, suite à une pétition dénonçant des synonymes sexistes et réclamant plus d'inclusivité, notamment pour les personnes appartenant à la communauté LGBT, l'éditeur a modifié ses entrées.

Désormais, selon la journaliste Ségolène Forgar, la définition ne renvoie plus uniquement à l'homme : la *femme* est devenue *l'épouse, la petite amie ou l'amante d'une personne*. C'est pourquoi, il est important de prendre en compte l'évolution sociétale et politique lorsque l'on s'intéresse à l'environnement distributionnel de la famille. Effectivement, avec ces évolutions, on peut imaginer que les mots apparaissant en cooccurrence avec *femme* peuvent désormais renvoyer à des termes de la famille du genre opposé à celui auquel on s'attendrait traditionnellement. Par exemple : « Marie est partie en voyage à Montpellier avec sa femme, qu'elle a rencontrée à l'université ». Cela illustre bien que la structure syntaxique reste identique, mais que les relations de genre impliquées par les référents évoluent, ce qui peut avoir un impact sur l'analyse linguistique des corpus contemporains.

Ces observations constituent la base théorique sur laquelle s'appuieront mes analyses dans les sections suivantes. Pour commencer, dans la prochaine partie, j'examinerai la manière dont ces différentes acceptions du mot *femme* se manifestent concrètement dans les dictionnaires et les corpus.

Partie 2. Analyse des régularités d'emploi du mot *femme*

Cette partie présente les différentes sources d'information mobilisées dans le cadre de la mise en place d'un système automatique de désambiguïsation du mot *femme*. Elle s'appuie d'abord sur les dictionnaires, qui constituent un premier point d'appui théorique pour comprendre l'organisation de la polysémie et pour relever des locutions utiles à la conception du système. Elle souligne toutefois les limites de ces ressources, qui ne rendent pas toujours compte de la diversité des usages en contexte.

La deuxième section propose une étude en corpus à partir de *Frantext*, en s'appuyant sur l'usage du concordancier dans une perspective exploratoire. Cette démarche m'a permis d'observer les occurrences du mot *femme* en contexte et d'identifier des régularités d'emploi susceptibles de contribuer à la désambiguïsation des deux sens principaux de *femme* : *être humain de sexe féminin* et *épouse*. À partir de ces observations, j'ai relevé des patrons récurrents, c'est-à-dire des configurations lexico-syntaxiques susceptibles d'être formalisées sous forme de règles. Cette étape correspond à une structuration des observations issues du corpus, en vue de leur exploitation ultérieure dans la conception du système automatique.

La troisième section complète cette analyse par l'utilisation de l'outil *Word Sketch Difference*, appliqué au corpus *frTenTen23* du *Sketch Engine* (Kilgarriff, 2004), afin de comparer les cooccurrents de *femme* et *épouse* sur un volume de données plus large. Cette approche m'a permis d'affiner les régularités identifiées en mettant en évidence des associations lexicales caractéristiques de chacun des sens, sur un volume de données plus large et un corpus plus diversifié que *Frantext*.

Finalement, l'ensemble de ces analyses permet de constituer une base de régularités lexico-syntaxiques présentée dans une quatrième section, issue à la fois des dictionnaires et d'observations en corpus, que je mobilise ensuite pour la conception du système automatique de désambiguïsation dans la troisième partie de ce mémoire.

2.1. La polysémie du mot *femme* dans les dictionnaires

Afin de commencer mes observations sur le mot *femme*, je me suis appuyée sur quatre dictionnaires : *Le Robert*, *Le Larousse*, le *dictionnaire de l'Académie française*, ainsi que le *Trésor de la Langue Française informatisé* (TLFi), tous dans leur version accessible en ligne. Ce choix s'appuie sur les travaux de Pruvost (2021), qui identifie ces dictionnaires comme des références majeures de la langue française, tant pour la qualité de leur élaboration lexicographique que pour leur rôle dans la description et l'évolution du français.

Comme nous l'avons vu dans la section 1.2.2, les sens proposés dans les dictionnaires ne sont pas classés de manière systématique (Kilgarriff, 1997). Ils dépendent des choix des lexicographes et varient d'un dictionnaire à l'autre. Ces quatre dictionnaires distinguent néanmoins les deux sens principaux de *femme* : *être humain de sexe féminin* et *épouse*. Pour le premier sens, il est parfois subdivisé en sous-sens, tandis que le second ne l'est jamais.

L'observation des définitions dans les dictionnaires m'a permis d'établir une liste de locutions comprenant le mot *femme*, qui a contribué à la désambiguïsation des sens dans mon système automatique. Une présentation détaillée de cette liste est proposée en annexe A. Elle comprend

l'ensemble des locutions relevées lors de mes observations, qu'elles proviennent des dictionnaires ou des corpus utilisés dans les sections suivantes.

Toutefois, cette liste peut parfois être contestée, car les sens ne correspondent pas toujours à ceux proposés par les dictionnaires (Kilgarriff, 1997). Par exemple, les dictionnaires classent *femme veuve* comme relevant du sens *femme_humain*. Cependant, certains cas posent problème du point de vue de l'identification automatique. Dans les exemples (1) et (2), *veuve* ne fait pas partie d'une locution nominale, mais fonctionne comme attribut du complément (*laisser sa femme veuve, sa femme devient veuve*). Le problème ne tient donc pas à la validité de la locution elle-même, mais à la difficulté de distinguer automatiquement ces structures de surface, qui peuvent être formellement proches mais syntaxiquement différentes. Ainsi, le nom *femme* suivi de l'adjectif *veuve* ne permettent pas à eux seuls de désambiguïser le sens de *femme*.

(1) *Il est décédé, laissant sa **femme veuve** à 25 ans. (exemple construit)*

(2) *À cause de la guerre, ses fils deviennent orphelins et sa **femme veuve**. (exemple construit)*

De même, la locution *bonne femme* est parfois classée par les dictionnaires comme du sens *femme_humain* (*Le Robert, le Dictionnaire de l'Académie Française*), mais d'autres fois, elle apparaît dans la catégorie *femme_épouse*, comme dans le TLFi, qui la classe dans les deux sens. Dans ce cas, c'est la présence d'un déterminant possessif précédant la locution qui permet de choisir le bon sens. Cette difficulté à classer certaines locutions est soulignée par Le Draoulec et Rebeyrolle (2025), qui relèvent dans le TLFi une tendance à orienter plusieurs emplois de *femme* vers le sens conjugal (*femme adultère, femme au foyer*). Selon elles, ces classements posent problème, car certaines de ces expressions restent indéterminées hors contexte et peuvent aussi être interprétées dans le sens *femme_humain*.

Ces exemples montrent qu'une structure contenant le mot *femme* ne suffit pas toujours à déterminer son sens. Le Draoulec et Péry-Woodley (2026) donnent notamment l'exemple de l'expression *faire de cette femme la sienne*, qui peut sembler renvoyer au sens *femme_épouse*, sans pour autant impliquer nécessairement le mariage ou le statut d'épouse. Elles citent l'exemple (3) où l'expression ne signifie pas que le personnage veut épouser la femme, elle traduit plutôt une logique d'appropriation, dans laquelle le personnage cherche à établir un rapport de possession avec la femme. Cet exemple montre que le sens de *femme* ne peut pas toujours être déterminé à partir d'une expression seule, il dépend aussi du contexte dans lequel le mot apparaît.

(3) « Depuis, une seule obsession l'envahit : **faire de cette femme la sienne**... Il la suit sur les réseaux sociaux et sur les parcours de ses courses, s'inscrit à ses cours d'entraînement sportif... Il n'aura de cesse de briser sa vie, son bonheur afin qu'elle soit à lui. » (*Babelio, commentaire à propos d'un roman de Peter James*)

Cela met en évidence les limites d'une approche fondée exclusivement sur les classifications proposées par les dictionnaires, qui simplifient la réalité des usages. Dans cette perspective, il apparaît nécessaire de compléter cette approche par une analyse en corpus, afin d'observer les emplois effectifs du mot en contexte. Cela permettra d'affiner les critères de désambiguïstation en s'appuyant sur les régularités attestées dans les usages et plus uniquement sur des définitions théoriques.

2.2. Frantext : recherche de patrons morphosyntaxiques

Afin d'observer les usages effectifs du mot *femme* en contexte, j'ai utilisé l'interface *Frantext*, car il s'agit d'une référence en langue française et d'une source fiable scientifiquement, bien que *Frantext* soit très limité en termes de variété de textes. Il couvre une période allant de 1950 à nos jours et se compose de 5 703 textes, majoritairement fictionnels (90 %) et non-fictionnels (10 %), soit un total d'environ 274 millions de mots en juillet 2025⁶.

J'ai recherché des patrons morphosyntaxiques à partir du corpus contemporain, qui contient 776 textes et 58 millions de mots, datant de 1980 à aujourd'hui. J'ai retenu ce corpus contemporain afin de trouver des usages actuels du mot *femme*. D'abord, j'ai interrogé la base de textes en utilisant la forme *femme* au singulier, puis, dans un second temps, le mot *femmes* au pluriel. Cette distinction m'a permis d'observer dans quelle mesure les environnements distributionnels du mot *femme* diffèrent selon le nombre (singulier/pluriel), ainsi que le caractère propre de certaines associations de mots à l'une ou l'autre forme.

La requête *femme* renvoie à 32 213 occurrences et 14 736 pour la version au pluriel sur 58 millions de mots. Afin de situer la fréquence du mot *femme*, on peut le comparer à des noms très fréquents du corpus, tels que *jour* et *temps*, ainsi qu'à des noms du même champ sémantique comme *homme* et *filles*. Le Tableau 2 propose une comparaison de ces termes sous leur forme lemmatisée, c'est-à-dire en regroupant les différentes formes fléchies d'un même mot sous une forme de référence, afin d'en faciliter la comparaison.

Lemme	Fréquence (par million)	Nombre d'occurrences
<i>Jour</i>	1234,54	71 631
<i>Temps</i>	1199,40	69 592
<i>Homme</i>	1009,28	58 561
<i>Femme</i>	807,34	46 844
<i>Fille</i>	523,49	30 374

Tableau 2 : Fréquence des lemmes *jour*, *temps*, *homme*, *femme* et *filles* dans le corpus *Frantext*.

Ces résultats montrent que *femme* est moins fréquent que des noms comme *jour* ou *temps*, mais plus fréquent que *filles* et relativement proche d'*homme*. Dans l'ensemble, cela indique que le mot *femme* présente une fréquence relativement élevée dans le corpus.

Pour observer les usages de *femme*, j'ai procédé en deux étapes : j'ai trié par contexte immédiat (mot précédent/suivant). J'ai repéré ces régularités de manière exploratoire, en m'appuyant sur leur fréquence apparente dans les concordances, sans recourir à la fonction de fréquence de *Frantext*, l'objectif étant avant tout d'identifier qualitativement les régularités les plus saillantes à partir de l'exploration du concordancier. Cette démarche permet de mettre en évidence les tendances principales, sans pour autant constituer une analyse exhaustive de l'ensemble des contextes possibles. La Figure 2 illustre l'interface de *Frantext* lors de l'utilisation du concordancier avec un tri à gauche et la mise en avant de la locution *bonne femme*.

⁶ Fabre, C. (2025). *Méthodologie : corpus et linguistique outillée*. [Notes de cours]. Département Sciences du langage, Université de Toulouse II.

Ensemble de données		Contexte					
Forme		100		Exporter		Statistiques	
				Vue		< < > >	
Texte		Contexte gauche ↓ 1	Pivot	Contexte droit	Actions		
1601	E699	fil, ça va ? – Écoute, il se sépare d'avec sa bonne	femme	! C'est pas plus mal pour lui ! – Tu métonnes, elle	🔍	✕	
1602	E678	ce cri de ralliement : «Tu nous aides à faire une bonne	femme	de neige, papa ?» Alexandrin sourit. Elles	🔍	✕	
1603	R062	vidé leurs vins rouges en sortent. Il y a là une bonne	femme	, commerçante ou concierge dans le coin. Elle – Ah là	🔍	✕	
1604	E608	ils meurent tous. –Je suis désolé, madame. La bonne	femme	pleurait. De grosses larmes roulaient le long de ses	🔍	✕	
1605	R847	plus rien... Qu'est-ce que c'est que cette bonne	femme	? Oh ! mais y en a deux : une grosse, une petite... Avant et	🔍	✕	
1606	E608	milieu de mon jardin? Markus garda le silence, la bonne	femme	poussa un sanglot douloureux et continua tant bien que	🔍	✕	
1607	E117	faire peur aux poules d'eau ? «Je r'semble une bonne	femme	», dis-tu, et il ne te reste plus, pauvrete, que tes	🔍	✕	
1608	E645	s'était affaissée sous son poids. C'était une bonne	femme	gigantesque. Haute et ronde, le visage décapé sans	🔍	✕	
1609	R064	mal. Madame Garcia faisait la cuisine, grosse bonne	femme	aux mains grasses qui s'épongeait continuellement le	🔍	✕	

Figure 2 : Aperçu du concordancier de l'interface Frantext.

On y observe la locution *bonne femme*. Les dictionnaires classent majoritairement cette locution dans le sens *femme_humain*, ce qui rejoint les observations formulées dans la section 2.1, comme l'illustre l'occurrence de la ligne 1603 de la Figure 2. Cependant, après l'examen de ces occurrences, les usages attestés révèlent qu'en présence d'un déterminant possessif (par exemple *sa*), elle peut également renvoyer au sens *femme_épouse* (ligne 1601 de la Figure 2). L'analyse en corpus m'a donc permis d'affiner mes listes de locutions afin de déterminer si certaines d'entre elles doivent être classées uniquement dans un sens ou si, comme *bonne femme*, les deux sens peuvent apparaître.

Méthodologiquement, j'ai parcouru l'ensemble des occurrences afin d'identifier les régularités utiles à la conception d'un programme automatique de désambiguïsation du mot *femme* dans une approche à base de règles. Lorsque de nouvelles régularités apparaissaient, j'ai analysé plus en détail les occurrences correspondantes. Cela m'a permis de dresser une liste de patrons morphosyntaxiques autour de *femme*, tout en prenant en compte les exceptions, qui font partie des usages. J'ai également réalisé le même type d'observations pour le mot *femmes* afin de m'assurer que le passage au pluriel n'apporte pas de régularités supplémentaires. Cependant, cette analyse n'a permis d'identifier que peu de patrons nouveaux par rapport à ceux déjà observés au singulier. De plus, ces patrons morphosyntaxiques sont principalement liés au sens *femme_humain*. Ils ne seront donc pas conservés dans le système automatique final, dans la mesure où celui-ci sera surtout construit autour des indices permettant d'identifier le sens *femme_épouse*, comme nous le verrons plus en détail par la suite.

Pour classer ces régularités dans un sens ou l'autre, je me suis appuyée sur mes premières observations dans les dictionnaires, sur mon intuition de francophone native, ainsi que sur le remplacement par un synonyme tel que *épouse* pour vérifier quel sens sélectionner pour chaque occurrence. Ensuite, j'ai parcouru le concordancier afin de chercher des contre-exemples. Cela m'a permis de faire un premier choix de classification.

Deux tableaux, regroupant les régularités observées pour le sens *femme_humain* et *femme_épouse*, sont disponibles en annexes B et C. Ces tableaux correspondent à une phase exploratoire de l'analyse, au cours de laquelle j'ai classé les régularités relevées au fil de l'observation des données. De ce fait, ils présentent un niveau de formalisation hétérogène : certains relèvent de schèmes morphosyntaxiques relativement généraux, comme *possessif + femme*, tandis que d'autres correspondent à des patrons lexico-syntaxiques plus spécifiques, construits autour de formes ou d'expressions récurrentes observées dans les corpus. De plus, dans l'optique d'éviter les répétitions, je n'y ai pas intégré les locutions, déjà disponibles en annexe A. Toutefois, mon classement présente quelques limites. Par exemple, j'ai classé

chaque femme dans le sens *femme_humain*, pourtant l'exemple (4) démontre que *chaque femme* peut également renvoyer au sens *femme_épouse*.

(4) *Chaque femme devrait prendre soin de son mari. (exemple construit)*

Toutefois, ces listes étaient simplement établies en suivant mes observations : elles ne prennent pas en compte un ordre de priorité entre les règles, ni les potentialités qu'offre un programme informatique, ainsi que ses limites. Elles ont ensuite été réorganisées pour la mise au point du programme de désambiguïsation, ce que nous verrons plus en détails dans la troisième partie de ce mémoire.

Ce travail m'a permis de constater que le sens *femme_épouse* présente moins de patrons morphosyntaxiques différents que le sens *femme_humain*. J'ai donc choisi de me concentrer sur l'identification de ce sens pour la suite de l'analyse, d'autant plus qu'il correspond à celui visé par la désambiguïsation dans les travaux de Tanguy et al. (2024, 2025). Ces listes seront complétées dans la section suivante par une analyse à l'aide de l'outil *Word Sketch Differences*.

2.3. *Word Sketch Difference : femme et épouse*

L'outil *Word Sketch Difference* du *Sketch Engine* permet d'observer les similitudes et les différences entre deux mots sémantiquement liés (Kilgarriff, 2004). Contrairement aux observations issues du concordancier, qui reposent principalement sur des séquences de surface, cet outil fournit des informations fondées sur des relations grammaticales et des constructions syntaxiques. L'objectif de cette analyse est de compléter la liste de patrons morphosyntaxiques, issue de *Frantext*, susceptibles de distinguer les deux sens principaux de *femme*, en comparant *femme* et *épouse*.

J'ai choisi de me concentrer sur les relations nominales dans la continuité des observations menées sur *Frantext*. Mon analyse porte donc principalement sur les modificateurs adjectivaux et nominaux du nom *femme*, qu'ils soient antéposés (*ADJ + femme*, *NOM + femme*) ou postposés (*femme + ADJ*, *femme + NOM*). J'ai également observé les structures en *femme de + N*, afin de compléter ma liste de locutions présentée en annexe A. Ce choix permet de privilégier les configurations directement liées au nom, plutôt que les relations verbales, souvent plus générales et moins directement exploitables pour cette analyse. En effet, à ce stade, il s'agissait de partir de patrons définis par la place des éléments autour de *femme* : les modificateurs adjectivaux et nominaux peuvent être décrits selon leur position antéposée ou postposée par exemple, tandis que les verbes ne se ramènent pas aussi facilement à une position fixe par rapport au nom.

Pour mener cette analyse, je m'appuie sur le corpus *frTenTen23* (Jakubíček et al., 2013). Il s'agit d'un corpus annoté automatiquement par *FreeLing* au niveau morphosyntaxique (*Part Of Speech*, POS). Contrairement au corpus contemporain de *Frantext*, relativement restreint (58 millions de mots), *frTenTen23* se distingue par son volume, avec 23,8 milliards de mots collectés entre 2022 et 2023. Les corpus *TenTen* sont constitués à partir de données issues du web, ce qui implique une grande diversité de types de contenus. Cette différence de taille permet d'examiner les régularités sur un ensemble de données bien plus vaste.

La requête *femme vs épouse* renvoie à environ 11 millions d'occurrences pour *femme* et 888 172 pour *épouse*. Cela correspond à une fréquence de 481 occurrences par million pour *femme* et 38 occurrences par million pour *épouse*. Cette valeur pour *femme* reste du même ordre de grandeur que celle observée dans *Frantext* (555 occurrences par million).

Comme le montre la Figure 3, les cooccurents sont présentés sous forme de tableaux distincts, en tenant compte des relations syntaxiques (Benzitoun et Cappeau, 2025). Chaque liste met en évidence les éléments caractéristiques de chaque mot, en faisant apparaître les cooccurents spécifiques à *femme* ou à *épouse*, ainsi que ceux qu'ils partagent, grâce à un code couleur et à leur fréquence d'utilisation. Les mots en vert sont spécifiques à *femme*, ceux en rouge à *épouse*, tandis que les éléments grisés sont partagés.

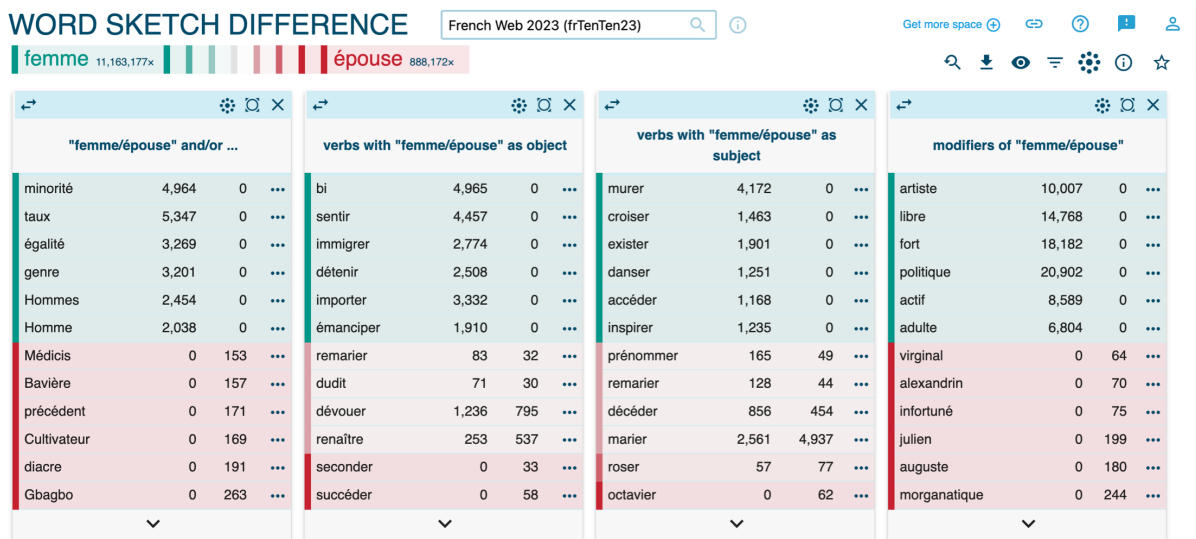


Figure 3 : Aperçu de l'interface de Word Sketch Difference avec la requête femme vs épouse.

Par exemple, dans la catégorie des *modifieurs*, certains adjectifs comme *libre*, *politique* ou *artiste* sont associés à *femme*, tandis que d'autres, tels que *morganatique*, apparaissent uniquement avec *épouse*. De même, dans les relations verbales, des verbes comme *marier* ou *remarier* sont davantage liés à *épouse*, ce qui suggère des contextes d'usage spécifiques. Dans cette analyse, je me limite aux cas où les cooccurents ne sont compatibles qu'avec un seul des deux mots, c'est-à-dire lorsque leur fréquence est nulle avec l'autre, afin de ne retenir que les contextes les plus discriminants.

Ces observations comportent cependant certaines limites. L'étiquetage des catégories grammaticales est parfois incorrect et certaines occurrences d'*épouse* renvoient en réalité au verbe *épouser*. De même, la lemmatisation automatique peut entraîner des rapprochements problématiques entre formes distinctes (*murer* et *mur* pour *femme mûre*). Enfin, certaines structures sont mal analysées sur le plan syntaxique, comme *remarié*, interprété comme une structure verbale avec objet, alors qu'il s'agit en réalité d'un participe (passif ou passé composé).

Par ailleurs, on observe que les cooccurents spécifiques à *épouse* relèvent souvent d'un registre plus soutenu ou de contextes historiques et littéraires, ce qui suggère un usage plus formel de ce terme par rapport à *femme*. Toutefois, bien que *femme* au sens *femme_épouse* et *épouse* soient des synonymes, ils ne sont pas interchangeables en contexte. Les usages d'*épouse* ne recouvrent donc pas nécessairement ceux de *femme_épouse*, ce que confirment Victorri et Fuchs (1996), pour qui la *synonymie pure* est *exceptionnelle* : remplacer une expression par une autre entraîne presque toujours une légère différence de sens. De plus, la fréquence des cooccurents spécifiques à *épouse* est très faible. Ces limites m'ont poussée à me concentrer sur le mot *femme* dans *Word Sketch Differences* et à identifier des usages propres au sens *femme_humain*, susceptibles de m'aider à désambiguïser les deux sens dans la suite de ce travail.

Finalement, cette analyse sur *Word Sketch Differences* m'a permis d'identifier plusieurs séquences de type *NOM + ADJ* associées au mot *femme*, qui semblent recouvrir uniquement le sens *femme_humain*. On peut citer par exemple *femme politique*, *femme indépendante* ou *femme migrante*. Cela complète mes précédentes observations dans les dictionnaires et sur *Frantext*.

2.4. Synthèse des régularités observées

Cette section propose une synthèse des locutions et des patrons lexico-syntaxiques identifiés précédemment. Elle vise à regrouper ces régularités afin de mettre en évidence les configurations les plus discriminantes pour la désambiguïsation des deux sens principaux du mot *femme*, et à préparer leur formalisation dans le système automatique présenté dans la partie suivante.

Pour le sens *femme_humain*, on relève des structures où *femme* est opposé à *homme*, comme *homme et femme* ou *ni homme ni femme*, mais également des séquences où le mot est répété ou mis en relation avec lui-même, comme *de femme en femme* ou *femmes entre elles*. S'ajoutent des déterminants démonstratifs (*cette femme*) et indéfinis (*une, chaque*). On observe aussi des constructions quantifiantes, où *femme* est précédé de quantifieurs de pluralité comme *nombreuses*, *quelques* ou *peu de femmes*, ainsi que des constructions comparatives du type *plus femme* ou *moins femme*.

Un autre ensemble concerne les emplois prédicatifs, où *femme* apparaît comme attribut de verbes comme *devenir*, *être*, *naître* ou *rester femme*, ainsi que des constructions telles que *en tant que femme*. Des adjectifs qualificatifs apparaissent aussi bien avant qu'après le mot, comme *belle femme* ou *femme âgée*. Par ailleurs, *femme* apparaît dans des compléments du nom de type *nom + de + femme* (*corps de femme*, *voix de femme*) ou *nom + pour + femme*. Enfin, on note des cas où *femme* est l'objet d'une transformation, avec des verbes comme *déguiser*, *habiller en femme* ou *se sentir femme*. Les dictionnaires attestent aussi des locutions figées, comme *femme de ménage* ou *femme au foyer*, qui constituent des indices du sens *humain*.

Pour le sens *femme_épouse*, les premiers indices concernent la présence d'un déterminant possessif, qu'il soit simple (*ma, ta, sa, votre, leur femme*) ou accompagné d'un adjectif (*première, future femme*). Les structures relationnelles en *de* constituent un autre indice, comme *la femme de + nom propre* ou *femme d'un + nom de métier*. Du côté des configurations verbales, *femme* apparaît avec des verbes d'acquisition ou de rupture comme *prendre*, *quitter* ou *choisir pour femme*, ainsi qu'avec le verbe *avoir* dans *avoir femme et enfants*. On retrouve également des coordinations avec *mari* (*mari et femme*, *mari-femme*) ou avec *enfants*, qui constituent des indices forts du sens conjugal. Enfin, *femme* apparaît dans des constructions relatives du type *nom + dont la femme* (*un homme dont la femme*), ainsi qu'avec d'adjectifs à valeur conjugale comme *chérie*, ou de termes plus institutionnels comme *morganatique* ou *légitime*.

À ce stade, il ne s'agit pas encore de déterminer précisément quels patrons seront conservés comme règles, mais de dresser un état des configurations récurrentes associées aux deux sens. Nous avons vu que, pour le sens *femme_humain*, les patrons relevés sont nombreux, variés, et parfois difficiles à regrouper dans des règles fixes. À l'inverse, les indices associés au sens *femme_épouse* semblent plus limités et plus stables, notamment la présence d'un déterminant possessif ou certaines constructions relationnelles. Cette différence explique que la formalisation nécessite ensuite un travail de sélection et de hiérarchisation des régularités relevées.

Par ailleurs, ce classement repose sur des choix d'annotation, effectués en fonction du sens qui apparaissait le plus fréquent d'après mes premières observations en corpus. Certains emplois peuvent rester discutables, dans la mesure où la désambiguïisation lexicale ne donne pas toujours lieu à un accord total entre annotateurs humains. De même, certains exemples peuvent relever des deux sens selon le contexte. Le Draoulec et Rebeyrolle (2025) montrent par exemple que *femme au foyer* peut désigner aussi bien une *femme* dans son rôle social qu'une *épouse*. Toutefois, j'ai classé cette locution dans le sens *femme_humain*, afin de conserver un classement stable et exploitable pour la formalisation automatique.

Ainsi, ces résultats permettent de dégager une base descriptive exploitable pour la désambiguïisation. La partie suivante présentera la manière dont certaines de ces régularités ont été sélectionnées, hiérarchisées, puis formalisées sous forme de règles linguistiques dans le système automatique.

Partie 3. Élaboration d'un système automatique de désambiguïsation

Cette partie présente la conception d'un système automatique de désambiguïsation du mot *femme*, fondé sur les patrons lexico-syntaxiques identifiés précédemment. La première section présente la méthodologie utilisée pour créer un programme Python. Ce dernier repose sur la bibliothèque SpaCy, permettant l'analyse d'un texte et son traitement linguistique automatique (tokenisation, étiquetage morphosyntaxique, lemmatisation, analyse en dépendances, extraction d'entités nommées). Elle permet d'exploiter ces informations en parcourant les unités du texte et leurs propriétés⁷.

Une fois la méthodologie définie, les observations issues de l'analyse en corpus et des dictionnaires sont converties en un ensemble de règles linguistiques explicites. Ces règles font ensuite l'objet d'une formalisation et d'une implémentation dans un programme Python, présentées dans la seconde section. L'objectif est de transformer les régularités identifiées en conditions applicables automatiquement au texte. Cette étape s'accompagne de tests réguliers sur un échantillon de 300 occurrences du corpus contemporain de *Frantext*, qui permettent à la fois d'ajuster progressivement les règles, d'optimiser leur ordre d'application et d'identifier les principales limites du système, ainsi que les ajustements mis en place pour y remédier.

La quatrième section propose une annotation de cas limites par trois annotateurs. L'objectif est de trancher les derniers cas pour lesquels des hésitations persistaient, notamment lorsqu'il s'agissait de classer certaines locutions dans un sens ou dans l'autre, comme *femme fidèle* ou *femme infidèle*. Cette étape m'a ensuite permis d'ajuster le programme à partir de décisions plus objectives.

Enfin, la dernière section vise à analyser la fréquence d'activation des différentes règles contenues dans le programme Python, afin d'en comprendre mieux le fonctionnement et leur fréquence respective dans la prise de décision. Cette analyse repose également sur l'échantillon de test de 300 occurrences issu du corpus contemporain de *Frantext*.

3.1. Méthodologie de conception du système automatique

Cette section présente la méthodologie que j'ai utilisée pour concevoir un système automatique de désambiguïsation du mot *femme*. Elle expose les principes généraux et les choix méthodologiques retenus, sans présenter de manière exhaustive l'ensemble des règles mises en place, qui seront présentées dans la section suivante.

L'analyse menée dans la partie 2 m'a permis de mettre en évidence un certain nombre de régularités lexico-syntaxiques relativement stables, qui semblent se prêter à une formalisation explicite. C'est pourquoi j'ai fait le choix d'une approche fondée sur des règles linguistiques. Cela s'inscrit dans une approche *knowledge-based* au sens de Navigli (2009), dans la mesure où elle repose sur des règles linguistiques explicites plutôt que sur un apprentissage supervisé à partir de données annotées. Par ailleurs, ce choix me permet de conserver un contrôle précis sur les critères de désambiguïsation, ainsi qu'une bonne interprétabilité des résultats, chaque décision pouvant être directement reliée à une règle linguistique identifiée.

⁷ Tanguy, L. (2026). Informatique pour le Traitement Automatique des Langues 2. [Notes de cours]. Département Sciences du langage, Université de Toulouse II.

Pour traduire ces règles linguistiques en code informatique, j'ai utilisé un programme Python reposant sur la bibliothèque SpaCy. Par exemple, pour la phrase suivante : *Le petit chat est mort*, SpaCy fournit pour chaque token des informations telles que sa catégorie grammaticale et son lemme, comme illustré dans l'exemple (5).

(5)

<i>Le</i>	<i>DET</i>	<i>le</i>	
<i>petit</i>	<i>ADJ</i>	<i>petit</i>	
<i>chat</i>	<i>NOUN</i>		<i>chat</i>
<i>est</i>	<i>AUX</i>	<i>être</i>	
<i>mort</i>	<i>ADJ</i>	<i>mort</i>	

SpaCy me permet donc d'utiliser ces informations pour formuler des règles reposant, principalement sur des patrons de surface, définis à partir de séquences de lemmes et de catégories grammaticales (par exemple *femme de + nom*). Dans certains cas plus spécifiques, j'ai également recours à des informations issues de l'analyse en dépendances, notamment pour identifier la présence d'un déterminant possessif, ainsi qu'à la reconnaissance d'entités nommées afin de distinguer les compléments désignant des personnes de ceux renvoyant à des lieux ou à des objets. Le système combine ainsi des indices de surface et, plus ponctuellement, des informations syntaxiques et sémantiques fournies par SpaCy.

J'ai choisi de travailler principalement sur les lemmes plutôt que sur les formes de surface, afin d'éviter les problèmes liés aux variations morphologiques (par exemple *femme / femmes*). De plus, nous avons vu dans la partie 2 que le singulier et le pluriel n'apportent de patrons morphosyntaxiques différents, ce qui permet de regrouper ces formes sous un même traitement. J'ai également normalisé les formes en les mettant en minuscules, afin de ne pas être dépendante de la casse (par exemple *Femme / femme*). Cela permet de simplifier les conditions et de rendre les règles plus générales.

Au fur et à mesure de la conception et de l'ajout de règles linguistiques dans mon programme, je l'ai testé sur un échantillon de 300 occurrences aléatoires de *femme* du corpus contemporain de Frantext, afin de vérifier le bon fonctionnement de mon programme, ainsi que la pertinence des règles en contexte.

J'ai d'abord annoté cet échantillon manuellement, en me basant sur mes observations précédentes. Dans la mesure où les régularités utilisées pour la désambiguïsation ont été identifiées à partir de ce même corpus, les occurrences à annoter présentaient des configurations déjà observées, ce qui a facilité cette étape. Le contexte, étendu à 200 caractères, a également contribué à limiter les ambiguïtés. J'ai ensuite testé les règles de désambiguïsation de *femme* de mon programme sur ce corpus, ce qui m'a permis de valider progressivement les différentes règles et d'identifier certains cas problématiques.

3.2. Formalisation des règles linguistiques et difficultés identifiées

Cette section propose un parcours des différentes règles implémentées dans le programme, dont la liste complète figure en annexe D, ainsi que des principales limites rencontrées lors de l'implémentation et des tests. Comme nous l'avons vu dans la partie 2, les patrons morphosyntaxiques renvoyant au sens *femme_humain* sont très nombreux et plus difficiles à formaliser dans un système automatique. J'ai donc choisi de me concentrer sur la désambiguïsation du sens *femme_épouse*, qui regroupe moins de patrons différents à implémenter dans un système automatique. En me concentrant sur le sens *femme_épouse*, je contribue aussi à la désambiguïsation du sens *femme_humain*, puisque l'identification des occurrences

relevant du premier permet de mieux isoler celles du second, qui peuvent en outre être confirmées par certains indices propres.

Mon programme parcourt le texte token par token et, lorsqu'il rencontre une occurrence de *femme*, il teste successivement les différentes conditions définies à partir des observations linguistiques. L'ensemble du programme repose sur une succession de conditions organisées dans un ordre précis.

La première règle hiérarchiquement présente dans le programme concerne *ex-femme* et *mari-femme*. Il ne s'agit pas d'une priorité linguistique, mais d'un choix imposé par les limites de segmentation de SpaCy, qui m'ont conduite à traiter ces formes en premier, car il analyse mal les séquences avec des tirets. Ensuite, la règle principale de désambiguïsation entre les deux sens principaux du mot *femme* repose sur la présence d'un déterminant possessif précédant le nom (*ma, ta, sa, votre, leur, son, mon, ton*), comme l'illustre l'exemple (6), ce qui constitue un indice fort en faveur du sens *femme_épouse*.

(6) *Sa femme est malade, impotente. (Frantext)*

La structure composée d'un déterminant possessif suivi de *femme* peut être complétée par l'ajout d'un adjectif entre les deux mots, voire d'un adverbe et d'un adjectif (exemple (7)). J'ai donc choisi, pour cette règle, de repérer la présence d'un déterminant possessif en relation de dépendance avec *femme*, afin de prendre en compte en une seule fois l'ensemble de ces configurations.

(7) *Son ex femme est venue lui rendre visite. (exemple construit)*

Cependant, cette règle n'est pas toujours valable. Par exemple, *femme* est précédé de *sa* dans une phrase comme : *C'est sa femme de ménage*. C'est pourquoi j'ai placé, avant cette règle, une liste d'exceptions relevant du sens *femme_humain*, afin de respecter l'ordre de priorité du programme. Ainsi, *sa femme de ménage* est traitée avant la règle des déterminants possessifs, ce qui permet une désambiguïsation correcte.

Ensuite, j'ai inséré plusieurs règles de désambiguïsation associées au sens *femme_humain* : une seconde liste dans laquelle *femme* est suivie d'un nom ou d'un adjectif, comme *politique* ou *mariée*, puis les séquences *femme au foyer* et *maîtresse femme*, ainsi que les cas où *femme* est précédé de certains quantifieurs (*certaines femmes, peu de femmes, etc.*).

Toutefois, j'ai retiré des locutions de la liste initiale, présentées en annexe A. En effet, mes premières observations m'avaient poussée à classer chaque patron et chaque locution dans un seul sens afin d'avoir une meilleure vue d'ensemble avant la mise en place du programme. Cependant, les tests ont montré que certaines de ces expressions pouvaient apparaître dans l'un ou l'autre sens selon le contexte, ce qui m'a conduite à resserrer cette liste. Par exemple, une *femme* peut avoir pour trait de personnalité d'être *aimante*, sans que cela soit lié à une relation conjugale comme le montre l'exemple (8), où *femme* est employée dans ce sens.

(8) *Sophie Grégoire Trudeau est une femme aimante et inspirante, elle est mère de trois enfants. (frTenTen23)*

Ces locutions peuvent donc être désambiguïsées par le possessif lorsqu'elles relèvent du sens *femme_épouse* ; dans les autres cas, c'est la règle par défaut finale, que je présenterai plus loin, qui permet de les classer dans le sens *femme_humain*.

Dans le corpus *Frantext*, j'ai également repéré plusieurs verbes, qui permettent de désambiguïser *femme_épouse*. Il y avait notamment *prendre*, *quitter* et *choisir*, éventuellement suivis d'une préposition (*prendre femme*, *choisir pour femme*). *Avoir une femme* semblait également être un patron récurrent au sens *épouse*. Néanmoins, j'ai aussi rencontré des occurrences avec la structure *avoir une femme*, qui ne renvoyaient pas au sens *femme_épouse* comme l'exemple (9).

(9) *Il y avait une femme qui courait derrière un camion dans lequel les Allemands emmenaient l'homme qu'elle devait épouser le jour (Frantext)*

J'ai tenté de mettre en place une règle dans mon programme automatique pour ajouter cette structure, comme en recherchant les occurrences de *femme* en tant que complément d'objet direct du verbe *avoir*. Cependant, cela identifiait les occurrences comme celles de l'exemple (9) comme étant également du sens *femme_épouse*. Après plusieurs tentatives et plusieurs tests pour vérifier si c'était une structure très récurrente ou non, j'ai choisi de supprimer cette règle du programme. En effet, l'objectif est de limiter au maximum le bruit en sortie, quitte à accepter un certain degré de silence. Pour le même type de raison, j'ai également retiré les structures qui identifiaient *plus de femme* (*il n'avait plus de femme*) et *dont la femme* (*Jean, dont la femme était commerçante*) comme sens *femme_épouse*.

Également pour le sens *femme_épouse*, j'ai ajouté deux règles : une liste de mots suivants *femme* tels que *légitime*, *répudiée*, etc., et les mots *future* et *ex* précédent *femme*. Toutefois, malgré mes recherches dans les dictionnaires et les corpus, certaines locutions restaient difficiles à rattacher clairement à un seul sens. J'ai donc choisi de soumettre les cas limites persistants à une annotation externe, afin d'éclairer certaines décisions, comme nous le verrons dans la section suivante.

Une autre règle fréquemment observée dans les corpus repose sur la structure du type *femme de*, suivie soit d'un nom propre (10), soit d'un déterminant et d'un nom de métier (11), éventuellement accompagné d'un adjectif.

(10) *1979 : j'ai retrouvé la première, l'an dernier, lorsque la neige arrêta le trafic de cars que Martine et moi allions prendre, chaque lundi matin ; la femme de Pierre pour son bureau, moi pour revenir à Paris après le week-end à Chavosse. (Frantext)*

(11) *Je raconte la chose à la femme d'un grand chirurgien en retraite, dont le fils est interne des hôpitaux, une catholique pieuse... (Frantext)*

Cependant, durant les tests du programme sur mon échantillon de 300 occurrences de *femme* sur *Frantext*, j'ai également remarqué que la structure *femme de* suivi d'un nom était trop permissive. L'exemple (12) le souligne, avec un nom qui ne classe pas cette occurrence au sens *femme_épouse*.

(12) *Madame Clément, la « femme de lessive », s'active dans le mouillé et le savonneux, environnée de buée. (Frantext)*

Pour remédier à ce type de problématiques, j'ai utilisé les entités nommées, car SpaCy permet de vérifier si le complément introduit par la préposition *de* est un humain (donc un nom propre ayant le sens *épouse*) ou un objet ou lieu (sens *femme_humain*). Les exemples (13) et (14) sont respectivement analysés par SpaCy dans mon programme comme renvoyant au sens *femme_épouse* dans le premier cas, *Paul* étant

identifié comme une personne, et au sens *femme_humain* dans le second, *Paris* étant identifié comme un lieu.

(13) *Et vous, vous êtes Anne, c'est Anne, non, je me trompe, je n'ai pas très bien entendu, la femme de Paul, l'épouse ? Je me trompe. Je me suis trompé ? Anne, c'est cela ? Anne ou Catherine. Catherine ? (Frantext)*

(14) *Je suis certes la femme de Paris la plus heureuse aux yeux du monde, la plus malheureuse en réalité. (frTenTen23)*

L'une des autres limites de SpaCy est qu'il n'étiquette pas toujours correctement les noms propres. J'ai donc dû étendre les catégories à *Miscellaneous (divers)* de SpaCy, qui est une catégorie d'entités nommées regroupant certains cas d'entités non classées ailleurs. Cela limite grandement le problème des prénoms mal catégorisés, mais peut également apporter quelques faux positifs.

J'ai complété ces règles avec la constitution d'une liste pour repérer les noms de métiers, qui arrivent également dans une structure du type *femme de + nom*. Cependant, la liste n'est pas exhaustive, il est impossible de lister toutes les professions ou les désignations sociales. Il serait toutefois envisageable d'utiliser une ressource externe ou une API lexicale (*Application Programming Interface*) afin d'enrichir automatiquement la liste des métiers et des rôles sociaux. Ce type de solution impliquerait l'intégration de ressources externes et un traitement supplémentaire des données, ce qui dépasse le cadre de ce travail. J'ai privilégié une liste construite manuellement, cohérente avec l'approche fondée sur des règles linguistiques adoptée dans ce mémoire.

J'ai également relevé plusieurs structures coordonnées contenant *mari*, comme *mari et femme*, *le mari ou la femme* ou encore *mari / femme*. Pour les regrouper dans une même règle, j'ai choisi de repérer la présence du mot *mari* dans les trois mots précédant *femme*, afin de couvrir ces différentes possibilités. J'ai aussi intégré une règle portant sur les structures où *femme* est coordonné avec des termes renvoyant aux *enfants*, comme dans *femme et les gosses*, *femme et enfants* ou *femme et fils*. Enfin, la dernière règle correspond à la valeur attribuée par défaut à *femme* : l'occurrence est classée dans le sens *femme_humain* lorsqu'aucune des règles précédentes n'a permis de la désambiguïser.

Finalement, ces règles permettent de couvrir un grand nombre de cas, sans pour autant garantir une désambiguïstation parfaite dans toutes les configurations : la présence de *mari* dans les trois mots précédant *femme*, par exemple dans une structure comme *le mari de cette femme*, peut renvoyer au sens *femme_humain*. De plus, certains contextes restent ambigus ou difficiles à trancher automatiquement, ce que nous verrons dans la suite de ce mémoire.

Après la mise en place du programme et les premiers tests, il me restait néanmoins des doutes sur certains cas limites. J'ai donc choisi de faire annoter ces quelques cas ambigus persistants par plusieurs personnes afin d'avoir un avis plus objectif.

3.3. Annotation des cas limites

Lors de la mise en place des règles linguistiques dans le système de désambiguïstation, certains cas restaient incertains et nécessitaient un avis extérieur. C'était notamment le cas de locutions comme *femme infidèle* ou *femme bafouée*, pour lesquelles le corpus fournissait peu d'occurrences exploitables. Ce faible nombre d'exemples rendait plus difficile la formulation d'une règle stable. J'ai donc demandé

à trois personnes francophones natives d'annoter huit occurrences du nom *femme*, afin de m'aider à déterminer le sens auquel les rattacher. Ces occurrences étaient des extraits du corpus contemporain de *Frantext* quand c'était possible, les autres ont été construites pour les besoins de l'analyse.

Les annotateurs ont eu accès à un guide d'annotation, disponible en annexe E. Ils pouvaient attribuer les deux sens à une occurrence lorsqu'ils estimaient que le contexte était insuffisant pour trancher. Les exemples proposés sont volontairement courts, afin de placer les annotateurs dans des conditions proches de celles du programme, qui repose principalement sur le contexte immédiat.

Les résultats de cette annotation sont présentés dans le Tableau 3. Ce tableau regroupe les huit phrases exemples utilisées, ainsi que le nombre d'annotateurs ayant retenu chaque sens.

Phrases	<i>Humain</i>	<i>Épouse</i>
C'est une femme fidèle .	2	2
La défunte femme aurait senti la mort arriver.	3	1
C'est une femme bafouée .	3	
La femme et les enfants sont partis le week-end à Paris.	2	1
La femme et les gosses .	1	2
C'est une chaste femme .	3	
C'est une femme infidèle .		3
Oui, je devinai ce que pouvait ressentir une femme adultère .	1	2

Tableau 3 : Résultats de l'annotation des cas limites.

Les six locutions : *femme fidèle*, *femme infidèle*, *femme bafouée*, *défunte femme*, *chaste femme* et *femme adultère* sont généralement classées au sens *épouse* dans les dictionnaires et *Word Sketch*.

Pour *femme adultère*, mon hypothèse était que tous les annotateurs classeraient cette locution dans le sens *femme_épouse*, mais l'un d'entre eux l'a classée comme étant de l'autre sens. L'annotateur a fait un lien avec le cas de *femme veuve*, qui ne renvoie plus à une relation conjugale active. Toutefois, si on regarde la définition du dictionnaire *Le Robert en ligne*, *adultère* signifie : « Fait d'avoir volontairement des rapports sexuels avec une personne autre que son conjoint. », ce qui renvoie explicitement à la relation au conjoint. D'autant que l'annotateur explique lui-même que « la femme est possiblement encore l'épouse de quelqu'un » au moment de l'adultère. À ce titre, bien que la classification proposée par l'annotateur soit discutable, la justification fournie confirme en réalité le lien avec la relation de couple. Pour cette raison, j'ai choisi de maintenir cette locution dans le sens *femme_épouse*.

Il n'en est pas de même avec *femme fidèle*, les annotateurs suggèrent que la *femme* peut aussi être *fidèle à son travail* ou *à ses principes*, pas uniquement à son époux, ce qui rend l'occurrence ambiguë. De même, ils supposent que dans mon exemple, c'est la *femme* dans le sens *être humain* qui a un comportement *fidèle*.

Pour *femme infidèle*, les annotateurs sont tous d'avis que contrairement à *femme fidèle*, le terme est uniquement utilisé dans le sens relationnel et renvoie donc au sens *femme_épouse*. Pour le cas de *femme défunte*, ils ont été influencés par le déterminant défini *la* et cela les a incités à classer cette occurrence dans le sens *femme_humain*.

Femme bafouée et *chaste femme*, classées dans les dictionnaires comme relevant du sens *femme_épouse*, sont ici interprétées uniquement dans le sens *femme_humain*. Cela montre que ces expressions ne peuvent pas être systématiquement associées au sens *femme_épouse*. On retrouve ici une différence entre les distinctions proposées par les dictionnaires et l'interprétation des occurrences en contexte : les catégories lexicographiques stabilisent certains sens, mais ceux-ci ne correspondent pas toujours directement aux usages observés en corpus (Kilgarriff, 1997). Par ailleurs, les trois annotateurs ont tous entre 21 et 26 ans, ce qui peut également jouer un rôle dans l'interprétation de ces expressions, dont certaines relèvent d'un registre plus soutenu ou d'un usage aujourd'hui moins fréquent. Ce décalage générationnel peut ainsi expliquer une certaine hésitation ou une interprétation différente du sens attendu.

Concernant *la femme et les gosses*, et *la femme et les enfants*, les exemples étaient volontairement très courts et ambigus afin de placer les annotateurs dans la même position que le système automatique, qui se base sur le contexte très proche. J'ai décidé de garder ces structures comme relevant du sens *femme_épouse*, car dans le corpus de test utilisés dans la section précédente, ces structures sont principalement utilisées dans ce sens.

En conclusion, ces résultats montrent que la tâche reste difficile, en particulier lorsque le contexte est limité, ce qui peut conduire à des hésitations ou à des divergences d'interprétation. Ce constat rejoint les observations de Navigli (2026) : dans l'évaluation qu'il présente, un annotateur humain atteint 91,3 de F-mesure, cela montre que la performance humaine en WSD ne correspond pas nécessairement à une réussite totale. Les divergences observées entre les trois annotateurs ne sont donc pas surprenantes, puisqu'elles concernent précisément des cas limites où les indices contextuels sont peu nombreux.

Cette annotation m'a toutefois permis de mieux identifier les occurrences les plus ambiguës et de trancher certains choix pour la suite de l'analyse. J'ai retiré *femme fidèle*, *bafouée*, *défunte* et *chaste* de mes listes afin de limiter les faux positifs en sortie de mon programme. Ce choix fait suite aux différents tests réalisés sur corpus, qui ont montré que certaines règles étaient trop permissives ou ne correspondaient pas toujours aux usages réels. De plus, les termes *bafouée*, *défunte* et *chaste* sont peu fréquents dans les données, ainsi leur exclusion a un impact limité sur les performances globales du système. L'élaboration du système s'est donc faite progressivement, en ajustant les règles à partir des résultats obtenus et des difficultés rencontrées. L'annotation des cas limites m'a également permis de confirmer certains choix et de mieux comprendre les situations les plus ambiguës.

Une fois le système stabilisé à la suite des ajustements et de l'annotation des cas limites, j'ai procédé à une analyse de l'application des règles sur le corpus de test. Cette étape permet d'observer le comportement du système final et d'identifier les règles les plus mobilisées dans la désambiguïsation.

3.4. Fréquence d'apparition des règles dans le système final

Afin de calculer la fréquence d'application des règles de mon programme, j'ai utilisé le même corpus de test, constitué d'un échantillon de 300 occurrences du mot *femme* issu du corpus contemporain de Frantext.

Pour analyser la répartition des règles, je ne les ai pas présentées selon chaque condition utilisée dans le programme, ce qui aurait produit un classement trop détaillé et peu lisible. J'ai choisi de les regrouper en catégories plus générales, correspondant aux types de structures linguistiques repérées dans les occurrences. Ce regroupement permet de faire apparaître les grands principes sur lesquels reposent les

règles, par exemple la présence d'un verbe associé au sens *femme_épouse* ou celle d'une locution explicitement conjugale. Ainsi, les structures impliquant des verbes comme *prendre femme* ou *quitter femme* ont été rassemblées dans la catégorie *Épouse avec verbe*. De même, les adjectifs ou constructions marquant explicitement une relation conjugale, comme *femme adultère* ou *femme légitime*, ont été regroupés dans la catégorie *Locutions épouse*.

J'ai également réuni les cas relevant du sens *humain* en catégories spécifiques : j'ai intégré les différentes expressions figées (*femme de ménage*, *femme de chambre*, etc.) dans *Exceptions humain*, et j'ai regroupé les contextes impliquant des quantifieurs (*certaines femmes*, *beaucoup de femmes*) dans *Quantifieurs humain*, bien que ces cas soient traités par plusieurs règles distinctes dans le programme.

J'ai rassemblé les structures de type *femme de + nom* dans la catégorie *Femme de*, indépendamment des sous-règles utilisées pour leur traitement. Enfin, j'ai regroupé toutes les occurrences ne correspondant à aucune règle spécifique dans la catégorie *Autre (humain)*, qui correspond au cas par défaut du programme. Le Tableau 4 présente la fréquence d'application de chacune de ces catégories de règles.

Types d'occurrences	Fréquence absolue	Fréquence relative
Cas particuliers d' <i>épouse</i>	0	0,0 %
Exceptions <i>humain</i>	6	2,0 %
Possessifs	68	22,7 %
Quantifieurs <i>humain</i>	0	0,0 %
<i>Épouse avec verbe</i>	2	0,7 %
Locutions <i>épouse</i>	0	0,0 %
<i>Femme de</i>	2	0,7 %
<i>Mari/enfants</i>	0	0,0 %
<i>Autre (humain)</i>	222	74,0 %
Total	300	100 %

Tableau 4 : Fréquence d'apparition des règles linguistiques dans un échantillon de 300 occurrences de *femme* sur le corpus contemporain de *Frantext*.

Nous pouvons constater que sur 300 occurrences, le système repose principalement sur deux règles : 74,0 % pour la condition par défaut si aucune règle spécifique n'a été appliquée et 22,7 % pour la présence d'un déterminant possessif. La forte proportion de la catégorie *Autre (humain)* s'explique par le fait que le système a été conçu en priorité pour détecter le sens *femme_épouse*. En l'absence d'indices linguistiques pour ce sens, les occurrences sont classées dans le sens *femme_humain*, qui joue ici le rôle de catégorie par défaut. À l'inverse, la fréquence relativement élevée des possessifs confirme qu'il s'agit d'un indice particulièrement fiable pour identifier le sens *femme_épouse*, ce qui valide le choix de lui accorder une place centrale dans le système. Les autres règles sont très peu utilisées, ce qui montre que certaines structures repérées restent rares.

Cette analyse permet ainsi de mieux comprendre le fonctionnement global du système et la fréquence respective des différentes règles dans la prise de décision. Nous pouvons toutefois faire l'hypothèse que cette répartition dépend du genre de corpus analysé et pourrait varier selon les usages observés. Dans ce contexte, il devient nécessaire d'évaluer plus précisément les performances du système sur des données différentes de celles utilisées pour sa conception et ses tests. La partie suivante présente la méthodologie d'évaluation, en particulier la constitution d'un *gold standard* permettant d'évaluer le système sur un corpus distinct du corpus contemporain de *Frantext*, utilisé jusqu'ici comme corpus d'observation et de test.

Partie 4. Évaluation du système de désambiguïsation

Cette partie présente la constitution d'un *gold standard* à partir du corpus parallèle OPUS (Lison et Tiedemann, 2016), en s'appuyant sur la traduction comme méthode de désambiguïsation, puis l'évaluation du système automatique présenté dans la partie précédente et sa comparaison avec un modèle de type LLM.

La première section justifie d'abord le choix de l'anglais comme langue de désambiguïsation du mot *femme*, à partir d'une comparaison avec le japonais fondée sur l'évaluation de trois traducteurs (*Google Traduction*, *DeepL* et *ChatGPT*). Elle présente ensuite les différentes traductions anglaises de *femme* dans OPUS, puis la sélection d'un sous-corpus anglais-français issu d'*OpenSubtitles 2016*, dans lequel j'ai extrait 300 occurrences contenant *femme* en français et *woman/women* ou *wife/wives* en anglais. Le recours à un corpus parallèle permet ainsi de filtrer et d'annoter automatiquement les occurrences à partir de leurs équivalents de traduction.

La deuxième section revient sur les principales limites du corpus OPUS, notamment les problèmes d'alignement et certaines ambiguïtés ou erreurs de traduction, ce qui justifie le tri manuel des données avant l'évaluation. La troisième section présente ensuite l'évaluation du système sur le *gold* des 292 occurrences restantes après le tri. Puis, avec les métriques présentées dans la section 1.2.7, j'ai calculé les performances du programme et j'ai analysé les erreurs produites par le système afin d'identifier d'où proviennent les limites de celui-ci. Enfin, dans une dernière section, j'ai comparé les performances de mon programme avec celles d'un LLM, avant d'analyser plus précisément les différences dans les types d'erreurs produites.

4.1. Constitution d'un *gold standard* à partir du corpus *OPUS*

Nous avons vu dans la section 1.2.6 que la traduction en anglais constitue une méthode de désambiguïsation relativement robuste. Dans cette continuité, j'ai choisi d'utiliser le corpus parallèle anglais-français *OPUS* pour l'évaluation de mon système automatique de désambiguïsation, ce qui permet d'assigner automatiquement une étiquette de sens aux occurrences françaises de *femme*.

4.1.1. Choix d'un corpus parallèle anglais-français

Le recours à un corpus parallèle se justifie par le fait que les correspondances entre langues peuvent fournir des indices pertinents pour la désambiguïsation des mots polysémiques. Comme le montre Apidianaki (2006), l'analyse des cooccurrences et des équivalents de traduction dans un corpus bilingue permet d'identifier différents usages d'un mot et d'en inférer le sens en contexte. Dans cette perspective, la traduction peut ainsi être considérée comme un indice indirect du sens.

Quant au choix de la langue anglaise comme corpus parallèle, il repose sur une comparaison préliminaire entre plusieurs langues, notamment l'anglais et le japonais, qui distinguent toutes deux les deux sens du mot *femme* dans leur traduction et dont mon niveau de maîtrise est suffisant pour ce type d'analyse. En me basant sur les choix de Śmigielska (2023), j'ai évalué les traducteurs *DeepL* et *Google traduction*, ainsi que le LLM *ChatGPT*, au niveau de la qualité de leur traduction de *femme* en anglais et en japonais dans un sous-échantillon de 50 occurrences de mon corpus de test, présenté dans la section 3.1. Le Tableau 5 présente le nombre de traductions correctes pour les trois traducteurs et les deux langues.

L'anglais et le japonais apparaissent tous deux comme des langues pertinentes pour la désambiguïsation du mot *femme* à partir de la traduction. En effet, pour chacune d'elles, les traducteurs ont produit un total de quatre erreurs de désambiguïsation, ce qui les place à un niveau de performance équivalent. J'ai toutefois choisi de retenir l'anglais, en raison de la plus grande disponibilité et diversité des ressources.

	Anglais	Japonais
<i>Google Traduction</i>	50	49
<i>DeepL</i>	48	48
<i>ChatGPT 5.2</i>	48	49
Total	146	146

Tableau 5 : Évaluation de 3 traducteurs sur les langues anglaise et japonaise.

Il est intéressant de comparer les erreurs générées par ces différents traducteurs afin de vérifier si les différents outils produisent les mêmes erreurs. Pour l'anglais, parmi les 4 erreurs relevées, 2 proviennent de l'exemple (15). Dans ce cas, la liste semblait fonctionner comme une énumération : *le travail* (sa situation pro), *la femme* (sa partenaire), *la chienne* (son animal domestique). Le manque de contexte semble avoir orienté *DeepL* vers le mauvais sens. Quant à *ChatGPT*, il a surinterprété le contexte et a traduit par *work, the woman, the bitch*. L'avantage de cet outil est qu'il peut justifier son choix si on lui demande et son objectif était de « conserver la dureté, la fatigue et l'amertume du narrateur ». Cela renvoie aux travaux de Śmigielska (2023), qui a testé *Google Traduction*, *DeepL* et *ChatGPT* sur des phrases contenant le verbe *louer*. Elle a montré que *ChatGPT 3.5* et *ChatGPT 4* peuvent produire soit la meilleure traduction, soit une erreur inattendue comme celle que nous venons de voir.

(15) les oreilles. Les bruits de la nuit, celui du métro qui passait régulièrement. Et toujours : le travail, la femme, la chienne ... Il y a 88 jours, l'homme que j'aime m'a quittée. La scène s'est déroulée le 25 janvier 1985, à (Frantext)

Les 2 autres erreurs de l'anglais viennent de l'exemple (16). Les deux traducteurs ont traduit *woman* à la place de *wife*. Cependant, le contexte de la phrase précédente montre bien que nous sommes dans le registre de la famille et que l'attendu était bien *wife*. Les traducteurs semblent s'être contentés de regarder l'article indéfini *une*, et ont considéré qu'il s'agissait probablement d'une femme nouvellement mentionnée dans le texte et non l'épouse de quelqu'un.

(16) n'a plus rien à attendre de la société. Roland Agret a un père de quatre-vingt-deux ans et un fils de cinq ans. Une femme admirable, aussi, que nous connaissons et aimons tous. Comment après sept ans passés dans les maisons d'arrêt (Frantext)

En japonais, il y a également eu 4 erreurs de traduction du sens de *femme*. Les 2 premières viennent également de l'exemple (16), avec les mêmes difficultés que pour l'anglais. Ensuite, l'exemple (17) a été traduit par *épouse* (妻 *tsuma*) en japonais. Il semble que *Google Traduction* a interprété le contexte à cause du mot *époux* dans la phrase. Cependant, la femme est mise en opposition avec l'homme dans cette occurrence, pourtant en japonais, le traducteur a traduit *l'épouse* (妻 *tsuma*) en opposition avec *l'homme* 男 (*otoko*), ce qui perd l'effet initial du texte en français.

(17) tour. Il est certains moments où la bouche de l'époux est bien proche de l'oreille de l'épouse, des moments où la femme reprend le dessus et où l'homme le plus coriace s'adoucît et confie inconsciemment ses secrets à sa compagne. - (Frantext)

Enfin, la dernière erreur provient de *DeepL* et concerne l'exemple (18), où la métaphore *au cou de femme* est remplacée par 首力^カ細く (*kubi ga hosoku*), qui signifie *au cou fin*. Ce choix est pertinent dans une optique de traduction plus naturelle, mais il ne répond pas aux exigences de notre analyse, car il efface la dimension métaphorique et genrée de l'expression originale, ce qui pose problème dans le cadre de ce travail.

(18) portaient en marge des annotations à l'encre, tracées d'une plume fine, peut-être par le chasseur au cou de **femme** et à la tempe ensanglantée. Chez un bouquiniste, il acheta les quinze ou vingt tomes de *La Comédie humaine* (Frantext)

Ainsi, malgré des erreurs liées au manque de contexte ou à des choix interprétatifs, les résultats obtenus confirment la pertinence du recours à l'anglais comme langue de comparaison pour la désambiguïsation du mot *femme*.

4.1.2. Les différentes traductions de *femme* en anglais

La collection OpenSubtitles 2016 (Lison et Tiedemann, 2016) est un corpus parallèle multilingue construit à partir de sous-titres de films et de séries, couvrant plus de 60 langues et plusieurs milliards de phrases alignées. Les données sont prétraitées (segmentation, normalisation, correction d'erreurs) puis alignées entre langues à partir des informations temporelles des sous-titres. Dans ce travail, nous utilisons le sous-corpus anglais-français de cette collection, en sélectionnant les 245 515 paires de phrases dont la partie française contient le mot *femme*.

Afin d'utiliser le corpus *OPUS* comme *gold standard*, j'ai d'abord dû trier les occurrences du mot *femme* qu'il contient. Sur les 245 515 paires de phrases initiales, 41 032 ne sont pas traduites par *wife*, *wives*, *woman* ou *women*. J'ai donc analysé ces cas afin de relever les différentes traductions de *femme* en anglais et d'identifier les éventuelles erreurs contenues dans un corpus tel qu'*OPUS*.

Parmi ces occurrences, certaines ne contiennent aucun équivalent à *femme* et s'expliquent par des problèmes d'alignement ou de traduction comme le montrent les exemples (19) et (20). Le décalage peut provenir du contexte, où une partie de l'information est portée par les phrases précédentes. Il peut aussi s'expliquer par certaines traductions automatiques.

(19) *Here.*

*Tu ne connais rien aux **femmes**.* (OPUS)

(20) *Thirteen years of my life.*

*M. Privat. Jamais ma **femme** ne vous le laissera.* (OPUS)

Par ailleurs, un grand nombre d'occurrences sont traduites par des équivalents de *femme* que je n'ai pas retenus dans la construction du *gold*, afin de limiter la diversité des termes et de conserver un ensemble d'étiquettes cohérent et exploitable pour l'évaluation. Le Tableau 6 regroupe les traductions anglaises du mot *femme* ayant une fréquence d'au moins 50 occurrences dans le corpus, en dehors des quatre traductions retenues pour le *gold*.

Mots	Fréquences
<i>Lady, ladies, lady's, ladies</i>	11 065
<i>Girl, girls, girl's</i>	5 390
<i>Female, females</i>	4 170
<i>Maid, maids, maiden, chambermaid, housemaid</i>	1 294
<i>Mrs, ms</i>	951
<i>Person</i>	835
<i>Housewife, housewives</i>	654
<i>Midwife, midwives</i>	556
<i>Girlfriend, girlfriends</i>	517
<i>Housekeeper</i>	399
<i>Human</i>	372
<i>Chick, chicks</i>	220
<i>Businesswoman</i>	198
<i>Gal, gals</i>	192
<i>Broad, broads</i>	192
<i>Madame, madam</i>	147
<i>Feminine</i>	95
<i>Womanhood</i>	73
<i>Spouse</i>	73
<i>Companion</i>	52
Total	27 445

Tableau 6 : Traductions anglaises du mot femme ayant au moins une fréquence de 50 dans OPUS.

Ce tableau montre que *femme* peut être traduit par une variété de termes en anglais, au-delà des seuls *wife* et *woman*. Toutefois, dans ce tableau, je n'ai conservé que les termes qui sont majoritairement utilisés pour traduire *femme*, en excluant ceux qui, bien que pouvant occasionnellement correspondre à *femme*, sont plus fréquemment traduits autrement, comme *mistress*, généralement rendu par *maîtresse*. De même, je n'ai pas pris en compte les nationalités ni les prénoms, dont le recensement n'était pas pertinent dans le cadre de cette étude. Enfin, je suis consciente que certains termes, comme *person*, ne constituent pas des équivalents systématiques de *femme*, mais j'ai choisi d'en conserver quelques-uns afin de rendre compte de certains usages observés.

Ainsi, le mot *femme* présente de nombreux synonymes et équivalents en anglais, ce qui complique l'exploitation directe d'un corpus comme *OPUS* et justifie le recours à une étape de pré-traitement.

4.1.3. Méthodologie de constitution du *gold standard*

Afin de constituer un corpus d'évaluation, j'ai sélectionné un échantillon de 300 occurrences à partir du sous-corpus anglais-français de la collection OpenSubtitles 2016 (Lison et Tiedemann, 2016). Cet échantillon n'a pas été équilibré selon les deux sens étudiés, puisque le sens de chaque occurrence n'était pas encore déterminé au moment de la sélection. Afin d'éviter toute ambiguïté dans l'interprétation du mot *femme*, je n'ai retenu que les occurrences traduites par *woman/women* et *wife/wives*. De plus, j'ai uniquement sélectionné les paires de phrases contenant une seule occurrence de *femme* en français et une seule occurrence de la liste *woman/women/wife/wives* pour l'anglais. Ce traitement permet de distinguer et d'annoter automatiquement, en toute objectivité, les deux principaux sens de *femme* : être humain de sexe féminin (*woman/women*) et épouse (*wife/wives*).

Voici deux exemples d'occurrences provenant de ce corpus pour le sens *femme_humain* (21) et pour le sens *femme_épouse* (22) :

(21) *I'm a pregnant **woman**!*
*Je suis une **femme** enceinte ! (OPUS)*

(22) *His **wife** and his son.*
*Sa **femme** et son fils. (OPUS)*

Enfin, j'ai procédé à une vérification manuelle de l'ensemble des occurrences sélectionnées afin de m'assurer de la qualité de l'alignement entre les phrases françaises et anglaises, ce qui m'a également permis d'identifier certaines limites dans les données. La section suivante en présente les principaux aspects.

4.2. Limites du corpus *OPUS*

Les sous-titres utilisés dans *OPUS* proviennent de la plateforme OpenSubtitles, où ils sont majoritairement produits par des utilisateurs, puis mis à disposition sous forme de fichiers synchronisés avec le média. Une partie de ces données peut également résulter de processus semi-automatiques, notamment via reconnaissance optique de caractères (OCR), ce qui introduit un certain bruit dans les données (Lison et Tiedemann, 2016).

Par ailleurs, en raison des contraintes spécifiques au sous-titrage (temps d'affichage, longueur des segments), les correspondances entre les langues ne sont pas toujours strictement littérales, comme le montre l'exemple (23), où la traduction française s'éloigne sensiblement de l'énoncé source. De plus, les unités lexicales ne se correspondent pas nécessairement terme à terme : les termes simples et complexes ne se traduisent pas systématiquement par des expressions de même longueur, ce qui peut entraîner des décalages dans l'alignement (Morin et al., 2004). Cela peut ainsi limiter leur utilisation comme référence linguistique directe, bien que dans ce cas précis, cela ne pose pas de problème pour désambiguïser *femme* automatiquement.

(23) *Hey, the landlord and his **wife** sure are making themselves scarce today.*
*Que sont devenus notre cupide logeur et sa **femme** ?(OPUS)*

Par ailleurs, l'usage en anglais montre que *wife* peut parfois être remplacé par *woman*, comme dans l'exemple (24), ce qui introduit une ambiguïté supplémentaire et complique l'exploitation d'un corpus parallèle. Ce phénomène reste toutefois peu fréquent, puisqu'il ne concerne que 3 occurrences sur les 300 analysées.

(24) *Only that the **woman** has an affair with the doctor.*
*Juste que sa **femme** a une liaison avec le toubib. (OPUS)*

Enfin, la traduction apporte également des contre-sens, comme c'est le cas dans l'exemple (25) où on passe de *femmes commandantes* à *la femme d'un officier*.

(25) ***Women** commanders.*
*La **femme** d'un officier. (OPUS)*

Finalement, parmi les 300 occurrences retenues de *femme*, j'ai relevé 8 cas où la traduction ne correspondait pas à celle attendue ou introduisait une ambiguïté, comme dans l'exemple (24), ce qui représente environ 3 %. Ces occurrences sont donc exclues du *gold* afin d'évaluer les performances de mon programme dans des conditions plus fiables. Le *gold* comporte finalement 292 occurrences.

Ces exemples ont montré que les occurrences ne se correspondent pas toujours directement d'une langue à l'autre, ce qui limite leur exploitation brute. Néanmoins, le corpus *OPUS* reste pertinent, à condition d'opérer un tri préalable afin de ne conserver que les occurrences correctement alignées.

4.3. Évaluation des performances

Cette section est consacrée à l'évaluation des performances du système de désambiguïsation. J'ai d'abord appliqué mon programme Python sur mon corpus d'évaluation et j'ai ensuite comparé les différences entre les attendus (*gold*) et les sorties de mon programme.

En première partie de ce mémoire (1.2.7), nous avons vu les différentes métriques utilisées pour calculer les performances d'un système de désambiguïsation, présentées par Habert (2005) et Navigli (2009). Dans le cas présent, on observe 0 faux positifs et 10 faux négatifs. En calculant les scores avec les différentes métriques, j'obtiens 100 % de précision, 96,58 % de rappel et 98,26 % de F-mesure. Ces résultats sont élevés, même si la comparaison avec d'autres systèmes de WSD doit être nuancée. En effet, les tâches de désambiguïsation lexicale étudiées dans la littérature sont souvent plus complexes et impliquent un plus grand nombre de sens (Raganato et al., 2017). Dans le cadre de ce travail, la distinction entre deux sens principaux simplifie la tâche, ce qui peut expliquer les scores obtenus.

Ces scores peuvent aussi être interprétés à partir des travaux de Picoche (1989), qui explique que la majorité des informations nécessaires à la désambiguïsation est encodée dans des indices linguistiques locaux, sans nécessiter une modélisation contextuelle profonde. Les scores obtenus confirment donc la pertinence des règles mises en place. Toutefois, il est nécessaire de s'intéresser plus précisément aux erreurs produites par le système.

Comme indiqué précédemment, l'objectif était de limiter les faux positifs, en considérant que la tâche consiste à retenir principalement le sens *épouse*, ce qui implique d'accepter un certain nombre de faux négatifs. Comme attendu, on observe 6 erreurs liées à des structures du type *avoir une femme*, comme dans l'exemple (26).

(26) *J'ai déjà **une femme** à la maison. (OPUS)*

Nous avons vu dans la section 3.2 que cette structure était mal traitée par le système, et qu'il avait été préférable de supprimer la règle correspondante, qui était trop permissive. Afin de vérifier la pertinence de ce choix, j'ai comparé le nombre de faux positifs et de vrais positifs associés à cette structure si je rétablis cette règle dans le programme et que je l'évalue à nouveau. On observe ainsi 7 vrais négatifs (7 occurrences correctement classées dans le sens *femme_humain*), comme l'exemple (27) pour seulement 6 faux négatifs (6 occurrences au sens *femme_épouse*, classées dans le sens *femme_humain*). Ce déséquilibre confirme que la règle introduisait davantage de bruit que de bénéfices, ce qui justifie son exclusion du système final.

(27) *Il y avait une femme dans les montagnes. Elle savait que je venais pour la fleur.*
(OPUS)

Ensuite, l'exemple (28) est plus intéressant, car il souligne les limites d'un système automatique fondé uniquement sur des règles linguistiques. Dans cette occurrence, le groupe nominal *la femme* ne contient pas en lui-même d'indice explicite du sens *épouse*, mais il est interprété comme tel par association avec le cadre familial évoqué. Ce type de relation anaphorique, fondée sur le contexte discursif, est identifiable pour un humain, alors qu'un système reposant sur des règles simples ne le permet pas toujours.

(28) *C'est, euh, c'est l'histoire du mec... c'est le type hyper catho qui s'installe dans la famille et qui veut coucher avec la femme ?* (OPUS)

L'exemple (29) est assez similaire dans la mesure où l'interprétation repose fortement sur les cooccurrences. La présence des termes *enfant* et *famille* oriente ici vers le sens *femme_épouse*. On peut supposer qu'un LLM sera en mesure d'identifier le sens correct puisqu'il est capable d'exploiter ces cooccurrences.

(29) *Tout comme moi... femme, enfant, famille.* (OPUS)

Ensuite, l'exemple (30) montre une autre limite d'un système à base de règles linguistiques et de listes. Effectivement, dans la liste des métiers intégrée à mon programme Python, il n'est pas possible de représenter de manière exhaustive l'ensemble des professions ou des désignations sociales. Il serait envisageable d'utiliser une ressource externe ou une API lexicale afin d'enrichir automatiquement la liste des métiers et des rôles sociaux.

(30) *Femme de samourai : triste destinée.* (OPUS)

Enfin, l'exemple (31) présente les limites de SpaCy. Mon programme devrait être en mesure de désambiguïser cette occurrence très simplement, pourtant il l'a identifiée dans l'autre sens. Après avoir vérifié l'analyse morphologique de SpaCy, j'ai pu constater que *ta* a été analysé comme un déterminant démonstratif par SpaCy, ce qui explique l'erreur de classification de mon programme.

(31) - *Ta femme t'attend, non ?* (OPUS)

Afin de situer ces résultats, il est intéressant de les comparer à ceux obtenus à l'aide d'un modèle de type LLM, afin d'évaluer dans quelle mesure une approche par règles peut être complétée ou dépassée par une approche exploitant un contexte plus large.

4.4. Comparaison des performances avec un LLM

Pour comparer les performances de mon programme Python aux performances d'un LLM, j'ai utilisé un modèle de grande taille (environ 120 milliards de paramètres), accessible via Ollama sous l'identifiant *gpt-oss:120b-cloud*. Contrairement à un autre modèle exécuté en local, celui-ci tourne sur une infrastructure distante (dans le *cloud*), ce qui me permet de bénéficier de performances élevées sans avoir besoin d'un matériel coûteux ou spécialisé. J'ai fait ce choix afin d'éviter de multiplier les requêtes vers un modèle commercial comme ceux proposés par *OpenAI*, en privilégiant une approche plus pratique consistant à traiter mes données en une seule fois.

L'extrait ci-dessous correspond au prompt que j'ai utilisé pour interroger le modèle *gpt-oss* via Ollama. Ce prompt est appliqué à un texte fourni en entrée standard, où le modèle traite chaque phrase successivement et, lorsqu'il rencontre une occurrence du mot *femme*, il attribue l'un des deux sens attendus (sens 1 ou sens 2).

Tu es un expert en sémantique lexicale et en désambiguïsation de sens.

Analyse attentivement la phrase suivante :

Phrase : "{phrase}"

Ta tâche est de déterminer le sens du mot "femme" dans ce contexte.

Sens possibles :

- sens1 : être humain de sexe féminin

- sens2 : épouse

Instructions :

- Choisis un seul sens.

- Réponds uniquement par "sens1" ou "sens2".

- Ne donne aucune explication.

En résultat, ce LLM obtient 98,62 % de précision, 99,31 % de rappel et 98,96 % de F-mesure. Le Tableau 7 regroupe les résultats de mon programme, présentés dans la section précédente, ainsi que les résultats du LLM afin de les comparer.

	Faux positifs	Faux négatifs	Précision	Rappel	F-mesure
Programme	0	10	100 %	96,58 %	98,26 %
LLM	4	2	98,62 %	99,31 %	98,96 %

Tableau 7 : Comparaison des performances du système basé sur des règles linguistiques et un LLM.

Le système fondé sur des règles se distingue par une précision parfaite (100 %), ce qui est particulièrement intéressant dans la perspective des travaux de Tanguy et al. (2024, 2025), dans la mesure où il limite la présence de faux positifs. Son rappel est en revanche inférieur au LLM, ce qui traduit la présence de quelques faux négatifs, mais cet aspect reste moins problématique dans le cadre de cette étude. À l'inverse, le LLM présente des performances plus équilibrées entre précision et rappel, avec des erreurs réparties entre faux positifs et faux négatifs.

Ses performances peuvent s'expliquer par sa capacité à exploiter des représentations contextualisées, qui permettent d'adapter l'interprétation du mot en fonction de son environnement linguistique, contrairement à l'approche par règles qui repose sur des indices explicites et locaux (Apidianaki, 2023). Toutefois, l'écart reste limité, ce qui indique que le système fondé sur des règles atteint un niveau de performance comparable sur le corpus *OPUS*, tout en étant plus transparent, ce qui constitue un avantage important d'un point de vue éthique.

Comme dans la section précédente, il est pertinent d'analyser plus précisément les types d'erreurs produites par le LLM, notamment pour vérifier si elles sont similaires à celles du programme basé sur

les règles. Afin de comparer plus finement les prédictions des deux systèmes, j’ai croisé leurs résultats sur l’ensemble des occurrences. Le Tableau 8 présente cette comparaison.

Cela montre que les deux systèmes obtiennent le plus souvent le même résultat, avec 278 occurrences correctement classées dans les deux cas. On observe toutefois des différences : le programme produit 8 erreurs qui ne sont pas partagées par le LLM, tandis que le LLM en produit 4 que le programme ne fait pas. Enfin, 2 occurrences sont mal classées par les deux systèmes. Ces résultats montrent que, bien que globalement performants, les deux systèmes ne se trompent pas systématiquement sur les mêmes cas, ce qui suggère une certaine complémentarité entre l’approche fondée sur des règles linguistiques et celle reposant sur un modèle de type LLM.

	LLM correct	LLM erreur
Programme correct	278	4
Programme erreur	8	2

Tableau 8 : Comparaison des prédictions du programme et du LLM sur le corpus d’évaluation (n = 292).

On peut donc examiner plus en détail les erreurs produites par le LLM. D’abord, le LLM semble parfois interpréter le contexte, comme dans l’exemple (32).

(32) Alors tu t'es marié à cette **femme** ? (OPUS)

Il y a 4 occurrences de ce type sur 292 dans lesquelles le LLM identifie le mauvais sens, ce qui suggère un phénomène récurrent. Il semble donc que lorsque le LLM repère un mot tel que *mari*, *marié* ou *épouser*, il classe alors l’occurrence comme *femme_épouse*. Toutefois, au contraire, il n’a pas été en mesure de classer correctement l’exemple (29), évoqué dans la section précédente. Le contexte de cette occurrence de *femme* contenait pourtant les mots *famille* et *enfants*.

Pour l’exemple (33), le LLM a classé l’occurrence dans le sens *femme_humain*, tout comme le programme Python, qui s’est basé sur le peu de contexte environnant.

(33) Une **femme** à assumer et les gens sauront que c'est insensé. (OPUS)

Le LLM a identifié 2 faux négatifs et 4 faux positifs. Dans le cas du LLM et du programme Python basé sur des règles linguistiques, le manque de contexte peut entraîner des erreurs de classification, ce qui explique la faible différence de performance entre les deux approches. Toutefois, le type d’erreurs diffère entre les deux. Le programme Python ne peut pas détecter les structures du type *avoir une femme*, ce qui entraîne des faux négatifs. Tandis que le LLM semble interpréter le contexte lorsque des mots associés au champ sémantique de l’*épouse* sont présents.

Ces résultats rejoignent les observations de Navigli (2026), qui montrent que les LLM récents peuvent atteindre des performances comparables à celles de systèmes spécialisés de WSD. Toutefois, l’absence de faux positifs en sortie du programme est prometteuse pour identifier correctement les occurrences du sens *femme_épouse* et permettre la continuité des travaux de Tanguy et al. (2024, 2025). De plus, cela ouvre des perspectives intéressantes pour la suite de ce mémoire, notamment dans l’analyse des usages de *femme* à travers différents genres de corpus.

Néanmoins, les analyses précédentes ont mis en évidence certaines limites propres à l’approche par règles et aux outils utilisés, notamment SpaCy. Ces limites rejoignent celles déjà soulignées dans la

première partie de ce mémoire, où la désambiguïsation lexicale dépend fortement de la richesse du contexte (Navigli, 2009), mais le système se montre malgré tout globalement fiable.

Ces évaluations mettent en évidence la fiabilité du système proposé, tout en soulignant certaines limites liées aux règles, aux outils et au corpus utilisé. La partie suivante s'appuie sur ce système afin d'analyser les usages du mot *femme* en corpus, en proposant une étude plus fine de la répartition de ses sens selon différents genres de données.

Partie 5. Variations inter-corpus et évolution diachronique des usages de *femme*

Dans cette dernière partie, nous étudions les variations des usages de *femme* à partir du système de désambiguïsation mis en place dans les parties précédentes. Nous nous intéressons ensuite à l'évolution diachronique des usages de *femme* et d'*épouse* dans *Frantext*. Pour les premières observations, je vais utiliser six corpus différents, qui seront présentés plus en détail dans la première section : *Revue.org*, *Wikipédia*, *Le Monde*, *CoMeRe*, *Orfeo* et *Frantext*. *Frantext* et *Orfeo* ne correspondent pas à des corpus uniques, mais à des ressources donnant accès à plusieurs textes ou corpus. Toutefois, dans la suite de cette partie, je les désignerai comme des corpus afin d'harmoniser la terminologie avec les quatre autres ressources analysées. L'objectif est d'observer si la proportion des deux sens de *femme* varie selon les corpus et si les usages de *femme_épouse* et d'*épouse* évoluent selon les périodes étudiées.

Nous reprenons ici l'hypothèse formulée par Tanguy et al. (2024, 2025) selon laquelle le sens *femme_humain* domine dans les usages de *femme*. Il s'agira alors de vérifier si cette hypothèse se confirme dans l'ensemble des corpus étudiés. Par ailleurs, je formule l'hypothèse que cette domination varie selon les genres discursifs. Je m'attends à ce que cette domination soit particulièrement forte dans les revues scientifiques (*Revue.org*), puis un peu moins importante dans *Wikipédia*, notamment dans les articles biographiques. Pour *CoMeRe*, l'hypothèse est plus difficile à formuler en raison de la diversité des contextes représentés, mais je prévois tout de même une majorité d'emplois au sens *femme_humain*. Dans *Le Monde*, la répartition pourrait être un peu plus équilibrée, tandis que le corpus oral *Orfeo* pourrait présenter une proportion plus importante d'emplois au sens *femme_épouse*, car ce corpus est centré sur la vie quotidienne. Enfin, pour *Frantext*, mes premières observations me laissent penser que le sens *femme_humain* sera particulièrement élevé (80%).

En ce qui concerne l'évolution de *femme_épouse* et d'*épouse* en diachronie, je fais l'hypothèse que *femme_épouse* est de plus en plus utilisé par rapport à *épouse* pour désigner la conjointe.

La première section porte sur les variations inter-corpus. J'y décris d'abord les différents corpus utilisés pour ces analyses, puis la méthodologie d'extraction et de préparation des échantillons afin d'annoter automatiquement les occurrences de *femme*. J'explique ensuite de quelle manière j'ai vérifié les performances de mon système automatique sur un échantillon de 100 occurrences pour chacun des corpus. Enfin, je présente les résultats concernant la répartition des deux sens de *femme* calculés sur 3000 occurrences de chaque corpus, ainsi que la variation des règles de désambiguïsation selon les corpus.

La deuxième section porte sur l'évolution diachronique des usages de *femme* et d'*épouse* dans *Frantext*. J'y présente d'abord le corpus utilisé et les périodes retenues pour l'analyse, correspondant aux décennies 1980-1989, 1990-1999, 2000-2009, 2010-2019 et 2020-2026. J'explique ensuite la manière dont les sous-corpus ont été constitués à partir des lemmes *femme* et *épouse*, puis comment le programme de désambiguïsation a été appliqué aux occurrences de *femme* afin de distinguer les emplois *femme_humain* et *femme_épouse*. Enfin, je présente les résultats de cette analyse diachronique, afin de vérifier si les usages de *femme_humain*, de *femme_épouse* et d'*épouse* évoluent selon les périodes étudiées.

5.1. Variations des usages de *femme* selon les corpus

5.1.1. Présentation des corpus comparés

Pour cette étude, j'ai choisi de travailler sur différents types de corpus afin de comparer les variations des usages de *femme* selon les genres discursifs. J'ai sélectionné six corpus : *Revue.org*, *Wikipédia*, *Le Monde*, *Frantext*, *CoMeRe* et *Orfeo*. Le Tableau 9 présente leurs principales caractéristiques afin de mettre en évidence les différences entre ces corpus. Il permet de comparer leur genre discursif, leur modalité, le nombre d'occurrences de *femme* relevées dans chaque corpus, ainsi que leur taille totale. Le nombre brut d'occurrences de *femme* n'étant pas directement comparable d'un corpus à l'autre, j'ai calculé une fréquence pour 100 000 mots. Cette normalisation permet de comparer plus précisément la présence de *femme* dans les différentes ressources étudiées.

Nom du corpus	Genre	Modalité	Nombre d'occurrences de <i>femme</i>	Taille du corpus (en nombre de mots)	Fréquence pour 100 000 mots
<i>Revue.org</i>	Revue scientifique	Écrit	149 485	200 millions	74,74
<i>Wikipédia</i>	Articles encyclopédiques	Écrit	339 982	720 millions	47,22
<i>Le Monde</i>	Articles journalistes	Écrit	93 622	200 millions	46,81
<i>Frantext</i>	Textes fictionnels et non-fictionnels	Écrit	33 321	50 millions	66,64
<i>CoMeRe</i>	Échanges en ligne	Écrit	322	1,2 millions	26,83
<i>Orfeo</i>	Interactions orales de la vie quotidienne	Oral	963	2,9 millions	32,91

Tableau 9 : Caractéristiques des corpus utilisés pour l'étude des variations de *femme*.

J'ai analysé les occurrences de *femme* dans les 4 corpus écrits : *Revue.org*, *Wikipédia*, *Le Monde* et *Frantext* afin de faire suite aux travaux de Tanguy et al (2024). Il s'agit des corpus qu'ils ont utilisés dans leurs recherches et sur lesquels ils souhaitent poursuivre leur étude en intégrant le mot *femme*. Toutefois, contrairement à leur travail, qui repose sur le corpus *Frantext* du XXe siècle, j'ai choisi d'utiliser le corpus contemporain de *Frantext*, afin de rester dans la continuité de mes précédentes observations. J'ai complété ces corpus avec *CoMeRe*, qui contient des différents types d'échanges en ligne comme des tweets et des SMS. Ce corpus est moins volumineux que les autres ressources étudiées, mais il reste intéressant dans la mesure où il donne accès à des formes d'écrit plus spontanées.

Enfin, j'ai sélectionné *Orfeo* afin de comparer les usages écrits et oraux et de vérifier si la modalité influence la répartition des deux sens de *femme*. Ainsi, j'ai regroupé les occurrences de 12 corpus de la ressource *Orfeo*. J'ai exclu le corpus *FLEURON* (30 267 mots), car il comporte des interactions impliquant des locuteurs non natifs du français, ainsi que le corpus *French Oral Narrative Corpus* (131 794 mots), car il repose sur des productions narratives stylisées, qui ne correspondent pas à des usages spontanés en interaction. Ainsi, le corpus oral du *CEFC-Orfeo* compte initialement environ 3 millions de mots⁸, mais la taille du sous-corpus oral retenu peut être estimée à environ 2,9 millions de mots.

⁸ Fabre, C. (2024). Linguistique outillée : Exploration du Corpus d'étude du français contemporain (CEFC / projet ORFEO). [Diapositives de cours]. Département Sciences du langage, Université de Toulouse II.

5.1.2. Préparation des échantillons pour l'annotation automatique

Afin de mener cette analyse sur les usages, j'ai d'abord extrait toutes les occurrences du lemme *femme* des 6 différents corpus. J'ai sélectionné les phrases contenant une seule occurrence de *femme* afin de faciliter l'annotation automatique, où une occurrence correspond à une ligne et donc à un sens uniquement. Cela m'a également permis de constituer des échantillons aléatoires.

J'ai créé deux échantillons aléatoires : un échantillon de 3000 occurrences pour calculer la fréquence au sens *femme_humain* et *femme_épouse* et un échantillon de 100 occurrences pour vérifier manuellement les performances de mon système sur chacun des corpus, qui servira dans la prochaine section. Pour les corpus *CoMeRe* et *Orfeo*, regroupant moins de 3000 occurrences de *femme*, j'ai gardé l'intégralité des occurrences de *femme* ainsi qu'un échantillon de 100.

J'ai également procédé à certains traitements avant de passer mon programme sur les corpus. Pour le corpus *Frantext*, j'ai directement extrait 3000 occurrences du lemme *femme* depuis la plateforme. Cela apportait des occurrences comme celle de l'exemple (34).

(34) *incarnation de la maladie. Une femme chauve, c'est une <*femme*> malade, peu importe qu'elle ait un pull magnifique, des (Frantext)*

Le fait qu'une même phrase contienne plusieurs occurrences de la chaîne de caractères *femme* ne pose pas directement problème pour l'annotation automatique des sens. Cependant, cela peut créer des doublons dans les résultats. Par exemple, un même échantillon pourrait contenir à la fois l'exemple (34), où l'occurrence ciblée est indiquée entre chevrons, et l'exemple (35), où une autre occurrence de *femme* apparaît dans le même contexte. Ces doublons risqueraient alors de biaiser les résultats.

(35) *L'alopecie, c'est l'incarnation de la maladie. Une <*femme*> chauve, c'est une femme malade, peu importe qu'elle ait un (Frantext)*

Dans les fichiers extraits de *Frantext*, l'occurrence ciblée est signalée entre chevrons. Afin que le programme ne traite que cette occurrence et ne comptabilise pas les autres emplois de *femme* présents dans le même extrait, j'ai remplacé ces derniers par *homme*. L'occurrence placée entre chevrons restait donc inchangée, tandis que les occurrences non ciblées étaient neutralisées. Ce choix permettait d'éviter qu'un même contexte soit pris en compte plusieurs fois dans les résultats, tout en conservant à la place de *femme* un nom commun occupant une position syntaxique comparable dans la phrase.

Pour *Frantext*, j'ai également effectué une opération chercher-remplacer afin de supprimer les chevrons qui encadraient les occurrences ciblées. Ces marques pouvaient en effet perturber l'analyse de SpaCy, notamment la tokenisation et l'annotation syntaxique des phrases.

Enfin, j'ai eu recours à un nettoyage du corpus *CoMeRe*. En effet, il contenait 359 fois la chaîne de caractères *femme*, mais certaines d'entre elles étaient des mentions de comptes comme dans l'exemple (36). J'ai donc neutralisé ces formes manuellement, en les remplaçant également par *homme*, afin qu'elles ne soient pas prises en compte comme des occurrences de *femme*. En finalité, il restait 322 occurrences analysables par mon programme de désambiguïsation.

(36) *Retour en vidéo sur la rencontre du Président @fhollande avec 3 générations de femmes au @MDDfemmes <http://t.co/NTNWz0NHMJ> (CoMeRe)*

Toutefois, avant de procéder à la désambiguïisation automatique des échantillons de 3000 occurrences, ou de l'ensemble des occurrences pour *CoMeRe* et *Orfeo*, j'ai d'abord vérifié le fonctionnement de mon programme sur un échantillon de 100 occurrences pour chaque corpus.

5.1.3. Vérification des performances du système selon les corpus

Mon programme de désambiguïisation ayant été testé uniquement sur le corpus contemporain de *Frantext* et sur *OPUS*, j'avais besoin de vérifier qu'il fonctionnait également sur les autres types de corpus avant de calculer la fréquence des différents sens. Cette vérification permettait de m'assurer que les résultats obtenus ne seraient pas biaisés par une baisse de performance du système sur certains corpus.

J'ai donc annoté manuellement un échantillon de 100 occurrences aléatoires pour chacun des corpus, avant d'y appliquer mon programme de désambiguïisation automatique afin d'évaluer les performances de chacun des corpus. Les résultats de ces évaluations sont présentés dans le Tableau 10.

Corpus	Précision	Rappel	F-mesure
<i>Frantext</i>	100 %	98 %	98,99 %
<i>Orfeo</i>	98,98 %	97,98 %	98,45 %
<i>CoMeRe</i>	98 %	100 %	98,99 %
<i>Wikipédia</i>	97,96 %	97,96 %	97,96 %
<i>Le Monde</i>	97,94 %	96,94 %	97,44 %
<i>Revues.org</i>	97,94 %	96,94 %	97,43 %

Tableau 10 : Performances du système de désambiguïisation selon les corpus.

On observe une stabilité globale des performances du système selon les corpus avec une F-mesure variant entre 97,43 % et 98,99 %. Ces résultats montrent que les règles de désambiguïisation mises en place ne sont pas uniquement adaptées au corpus sur lequel elles ont été élaborées, mais qu'elles restent efficaces sur des données de genres discursifs différents.

De plus, en analysant manuellement les erreurs obtenues, j'ai constaté qu'elles correspondaient à des erreurs déjà observées lors de mes évaluations précédentes du système dans la section 4.3. Elles ne semblent donc pas propres à un corpus particulier.

On y retrouve des erreurs de classification contenant *avoir une femme*, comme l'illustre l'exemple (37). D'autres erreurs sont liées à la liste non exhaustive des métiers utilisée dans le programme, comme le montre l'exemple (38), même si *roi* ne correspond pas exactement à un métier.

(37) *Mais en fait, 'Alî voulait **avoir deux femmes**, et le Prophète ne voulait pas, alors ils se sont mis d'accord avec Bilâl, parce que Bilâl était vraiment digne de confiance.*
(*Revues.org*)

(38) *Pourtant, des rumeurs courent d'une possible alliance de **la femme du roi Ferdinand**, Marie-Caroline d'Autriche, avec les Français.* (*Wikipédia*)

Il y a également quelques ambiguïtés liées au manque de contexte pour un programme basé sur des règles, comme l'exemple (39), où l'expression *femme ou petites amies*, peut être difficile à interpréter automatiquement sans prise en compte plus large du contexte.

(39) *Candidatures instrumentalisées, téléguidées, femmes ou petites amies de, tripatouillages de ces " votes démocratiques de la base " à coups de cartes achetées par paquets de dix ou de cent, champ de bataille rêvé pour les courants dont la mort si souvent annoncée n'est jamais constatée : la " tambouille " interne n'a pas toujours été reluisante. (Le Monde)*

Ces évaluations m'ont permis de me rendre compte de quelques limites de mon programme comme le montre l'exemple (40), où *hommes, femmes et enfants* ne renvoient pas au sens *épouse*, mais bien à l'*être humain* en opposition à *hommes*.

(40) *Des dirigeants étudiants originaires de la minorité bouddhiste vivant dans les collines du sud-est du pays (les Chittagong Hill Tracts) ont affirmé que mille deux cents de leurs coreligionnaires _ hommes, femmes et enfants _ avaient été massacrés, le 10 avril dernier, par des colons musulmans et des forces de sécurité. (Le Monde)*

Enfin, l'exemple (41) montre une spécificité des tweets. Ici, il s'agit d'une mention, ce qui rend l'interprétation plus difficile pour un programme fondé sur des règles. En effet, ce type de structure peut perturber l'analyse syntaxique produite par SpaCy. Dans ce cas précis, le système a classé correctement l'occurrence au sens *femme_humain*, notamment grâce à la règle par défaut. Néanmoins, ce type d'exemple montre que les corpus numériques peuvent contenir des structures particulières, qui constituent une limite potentielle pour l'analyse automatique.

(41) *Les femmes de @MeMv avec @RHammedi mobilisées pour # http : // t. co / 4 Zz 8 JuJ 3 MX (CoMeRe)*

Ces résultats permettent donc de poursuivre l'analyse quantitative des deux sens de *femme* dans les différents corpus. En effet, les performances du système restent stables d'un corpus à l'autre, avec un taux moyen d'erreurs de classification automatique d'environ 3,5 %.

5.1.4. Répartition des deux sens de *femme* selon les corpus

Après avoir vérifié la fiabilité de mon système sur un échantillon de 100 occurrences pour chaque corpus, je suis passée à l'analyse quantitative de la répartition des deux sens de *femme*. Cette analyse repose sur un échantillon de 3000 occurrences pour les corpus les plus volumineux, et sur l'ensemble des occurrences disponibles pour *CoMeRe* et *Orfeo*. Le Tableau 11 présente les résultats de cette analyse en pourcentage afin de permettre la comparaison entre les corpus.

Corpus	Femme_humain	Femme_épouse
CoMeRe	97,83 %	2,17 %
Revues.org	94,73 %	5,27 %
Le Monde	86,3 %	13,7 %
Orfeo	79,23 %	20,77 %
Wikipédia	77,8 %	22,2 %
Frantext	72,8 %	27,2 %

Tableau 11 : Répartition des deux sens de *femme* selon les corpus.

Ces résultats vont globalement dans le sens de l'hypothèse selon laquelle le sens *femme_humain* est dominant dans l'ensemble des corpus. Sa proportion varie toutefois fortement, de 72,8 % dans *Frantext*

à 97,83 % dans *CoMeRe*. À l'inverse, le poids du sens *femme_épouse* diffère nettement selon les corpus : il ne représente que 2,17 % des occurrences dans *CoMeRe*, contre 27,2 % dans *Frantext*.

Les résultats obtenus pour *Revue.org* confirment également la forte domination attendue du sens *femme_humain*, mais dans une proportion encore plus élevée que prévu, puisqu'il atteint 94,73 %. Pour *Wikipédia*, la répartition observée est proche de celle envisagée, avec 77,8 % d'occurrences relevant du sens *femme_humain*. Dans *Le Monde*, ce sens est en revanche plus dominant que prévu et représente 86,3 % des occurrences. Enfin, *Frantext* se distingue par une proportion plus faible du sens *femme_humain* que celle qui avait été envisagée, avec 72,8 % des occurrences.

Enfin, je m'attendais à trouver moins d'occurrences du sens *femme_humain* dans *Orfeo*, corpus composé de données orales souvent issues d'échanges de la vie quotidienne. Les résultats restent néanmoins proches de ceux observés dans les autres corpus, avec 79,23 % d'occurrences relevant de ce sens. Cette comparaison ne permet toutefois pas d'attribuer cette répartition à la seule modalité orale, puisque *Orfeo* se distingue également des autres corpus par les genres discursifs et les situations de communication qu'il rassemble. Les écarts observés peuvent donc dépendre à la fois de la modalité et du type de discours représenté.

Le résultat le plus inattendu est le taux si élevé d'occurrences au sens *humain* pour le corpus *CoMeRe*. Je m'attendais à une répartition plus équilibrée, car ce corpus regroupe différents types d'écrits numériques, comme des blogs, des forums ou des tweets. Cependant, en analysant qualitativement les occurrences de ce corpus, j'ai remarqué une majorité de tweets politiques et militants, ce qui explique en partie la forte présence du sens *femme_humain*.

Par ailleurs, afin de vérifier la pertinence du regroupement des corpus oraux d'*Orfeo*, j'ai observé la répartition des deux sens de *femme* dans chacun des sous-corpus. Cette répartition est présentée dans le Tableau 12.

Corpus	Caractéristiques	<i>Femme_humain</i>	<i>Femme_épouse</i>
CFPB	Interviews en parlé de Bruxelles	62,50 %	37,50 %
CFPP	Interviews conversationnelles	83,46 %	16,54 %
CLAPI	Situations diverses en interaction	82,05 %	17,95 %
CORALROM	Discours spontané	78,90 %	21,10 %
CRFP	3 situations de parole	81,25 %	18,75 %
OFROM	Monologues et interactions	72,16 %	27,84 %
REUNIONDETRAVAIL	Réunions de travail	91,67 %	8,33 %
TCOF	Interactions diverses	77,33 %	22,67 %
TUFS	Entretiens multilingues	78,61 %	21,39 %
VALIBEL	Enregistrements vocaux	79,75 %	20,25 %

Tableau 12 : Répartition des deux sens de *femme* dans les sous-corpus oraux d'*Orfeo*.

Les proportions restent globalement stables d'un sous-corpus à l'autre. Les deux écarts les plus importants concernent *CFPB*, où la proportion du sens *femme_humain* est plus faible, et *REUNIONDETRAVAIL*, où elle est au contraire plus élevée. Toutefois, ces deux sous-corpus comportent très peu d'occurrences de *femme*, ce qui rend leurs pourcentages moins représentatifs. Il semble donc pertinent d'avoir regroupé ces données sous la catégorie générale des corpus oraux d'*Orfeo*.

Finalement, ces résultats vont dans le sens du choix de Tanguy et al. (2024, 2025) d'exclure le mot *femme* de leurs analyses. En effet, les occurrences relèvent très majoritairement du sens *femme_humain*, alors que leurs travaux portent sur le lexique de la parenté, pour lequel seul le sens *femme_épouse* est pertinent.

5.1.5. Variation des règles de désambiguïsation selon les corpus

Après avoir observé la répartition des deux sens de *femme* selon les corpus, il est possible d'examiner plus précisément quelles règles de désambiguïsation ont été mobilisées dans chacun d'eux. Cette analyse permet de voir si certaines règles apparaissent davantage dans certains types de corpus et d'expliquer certaines différences observées dans la répartition des sens. Le Tableau 13 regroupe la fréquence d'apparition des différentes règles de désambiguïsation implémentées dans mon programme Python. Pour rappel, j'ai regroupé certaines règles syntaxiquement proches, que j'ai présentées dans la section 3.4 de ce mémoire.

Règles	<i>Revue.org</i>	<i>Wikipédia</i>	<i>Le Monde</i>	<i>Frantext</i>	<i>CoMeRe</i>	<i>Orfeo</i>
<i>Épouse</i> cas particuliers	0.07 %	0.47 %	0.33 %	0.00 %	0.00 %	0.00 %
<i>Humain</i> exceptions	2.33 %	6.87 %	3.40 %	3.13 %	1.55 %	4.88 %
Possessifs	3.57 %	17.03 %	10.30 %	24.07 %	1.55 %	19.42 %
<i>Humains</i> avec quantifieurs	0.17 %	0.20 %	0.23 %	0.00 %	0.00 %	0.93 %
<i>Épouse</i> avec verbes	0.03 %	0.17 %	0.03 %	0.13 %	0.00 %	0.10 %
<i>Épouse</i> locutions	0.13 %	0.00 %	0.10 %	0.07 %	0.00 %	0.00 %
<i>Femme de</i>	0.50 %	3.57 %	1.87 %	2.40 %	0.62 %	1.14 %
<i>Épouse</i> (<i>mari/enfants</i>)	0.97 %	0.97 %	1.07 %	0.53 %	0.00 %	0.10 %
<i>Femme else</i>	92.23 %	70.73 %	82.67 %	69.67 %	96.27 %	73.42 %

Tableau 13 : Fréquence d'apparition des différentes règles de désambiguïsation dans les corpus.

La règle par défaut *femme else* (*femme_humain*) domine dans tous les corpus, surtout dans *CoMeRe* avec 96,27 % et *Revue.org* avec 92,23 %. À l'inverse, les possessifs constituent la règle la plus discriminante pour le sens *femme_épouse* : ils sont très présents dans *Frantext* (24,07 %), *Orfeo* (19,42 %) et *Wikipédia* (17,03 %), mais beaucoup moins dans *Revue.org* (3,57 %) et *CoMeRe* (1,55 %). Les règles liées aux verbes, aux locutions et aux cas particuliers restent très marginales dans tous les corpus. Elles sont utiles pour repérer un maximum d'occurrences, mais elles expliquent peu les grandes différences entre corpus.

Wikipédia se distingue par une fréquence élevée de la règle *Femme de* : 3,57 %, contre 2,40 % dans *Frantext* et 1,14 % dans *Orfeo*, alors que ces deux corpus présentent pourtant un taux assez proche d'occurrences au sens *femme_épouse*. Ce résultat rejoint l'hypothèse selon laquelle le sens *femme_épouse* pourrait être davantage représenté dans *Wikipédia* en raison des articles biographiques, où les conjointes de certaines personnalités peuvent être mentionnées comme le montre l'exemple (42).

(42) Elle est la **femme de Julien Bontemps**, double champion d'Europe et double champion du monde et tout comme elle, sélectionné olympique.

Il est également intéressant de relever que les deux corpus présentant le plus haut taux de *femme_humain*, *Revues.org* et *CoMeRe*, sont aussi ceux où les exceptions au sens *humain*, comme *femme de ménage* ou *femme de lettres*, sont les moins représentées. Cela suggère que les différences entre corpus ne se limitent pas à la répartition entre *femme_humain* et *femme_épouse*, mais concernent aussi les types d'emplois associés au sens *humain*.

Plus globalement, nous pouvons constater que chaque règle apparaît dans au moins un corpus, y compris en dehors de *Frantext*, qui a servi à élaborer et tester le système. Cela confirme l'intérêt d'avoir conservé des règles spécifiques, bien que leur utilité varie selon les corpus. Certaines règles sont peu fréquentes dans l'ensemble des données, mais permettent tout de même de repérer des emplois particuliers qui auraient pu échapper à une approche plus générale.

Toutefois, ce tableau montre une nouvelle fois que les deux principales règles de désambiguïsation de *femme* sont les possessifs et le classement par défaut au sens *femme_humain*. On peut alors se demander si un programme plus simple, fondé principalement sur la détection des possessifs et sur quelques exceptions au sens *humain*, comme *femme de ménage*, ne permettrait pas d'obtenir une précision plus élevée. Cependant, ce type d'approche aurait un rappel plus faible, puisqu'elle manquerait un certain nombre d'occurrences au sens *femme_épouse*.

Cette piste pourrait être intéressante dans le cadre des travaux de Tanguy et al. (2024, 2025), dans la mesure où *filles* constitue également une limite pour l'analyse distributionnelle du lexique de la parenté en raison de sa polysémie. Il serait possible d'envisager un système de désambiguïsation plus simple que celui proposé ici pour *femme*, reposant seulement sur les règles fortement discriminantes : les possessifs pour repérer les emplois parentaux, des exceptions lexicales comme *filles au pair*, puis un classement par défaut des autres occurrences dans le sens d'*enfant de sexe féminin*. Une telle solution ne permettrait pas de repérer toutes les occurrences pertinentes, mais elle pourrait constituer une méthode temporaire pour réintégrer *filles* dans une analyse distributionnelle sans passer par une longue étude de corpus.

Cette première analyse a donc permis de montrer que les usages de *femme* varient selon les corpus, aussi bien dans la répartition des deux sens que dans les règles de désambiguïsation mobilisées. Après avoir observé ces variations selon les genres discursifs, il s'agit maintenant d'examiner si les usages de *femme* évoluent également dans le temps.

5.2. Évolution diachronique des usages de *femme* et d'*épouse* dans *Frantext*

5.2.1. Présentation du corpus et des périodes étudiées

Cette section est consacrée à l'observation de l'évolution diachronique des usages de *femme_humain*, de *femme_épouse* et d'*épouse*, afin de voir si la répartition entre ces formes varie selon les périodes étudiées. L'objectif est notamment de vérifier si l'emploi d'*épouse* diminue progressivement au profit de *femme_épouse*. Il s'agit également d'observer si la proportion respective des deux sens de *femme* évolue au cours du temps et si l'un d'eux tend à devenir plus fréquent que l'autre.

Pour cette analyse, j'ai choisi d'utiliser une nouvelle fois le corpus contemporain de *Frantext*. Ce choix s'explique par deux raisons. D'une part, ce corpus permet d'observer des usages récents, sur une période allant des années 1980 à aujourd'hui, comme je l'ai déjà expliqué dans la section 2.2, consacrée à

Frantext. D'autre part, la plateforme *Frantext* permet d'effectuer des tris avant l'exportation, ce qui m'a permis de créer 5 sous-corpus correspondant aux périodes étudiées.

J'ai donc trié les données selon les périodes suivantes : 1980-1989, 1990-1999, 2000-2009, 2010-2019 et 2020-2026. Ce découpage par décennies permet d'observer l'évolution des usages sur l'ensemble de la période couverte par le corpus contemporain de *Frantext*. J'ai ainsi extrait 5 sous-corpus contenant le lemme *femme*, puis autres contenant le lemme *épouse*, chacun correspondant à l'une des périodes étudiées.

Ensuite, comme pour la section précédente, j'ai remplacé les doublons de *femme* par *homme* pour ne comptabiliser qu'une fois chaque apparition du mot *femme*. Puis, j'ai passé mon programme de désambiguïsation sur chaque sous-corpus et à l'aide d'une commande *egrep*, j'ai compté le nombre d'occurrences de *femme_humain* et *femme_épouse* sur mon fichier de sorties. Pour les occurrences d'*épouse*, il me suffisait de relever le nombre d'occurrences directement sur *Frantext*.

5.2.2. Résultats de l'analyse diachronique

Les effectifs bruts obtenus sont présentés en annexe F. Ils ont servi de base aux calculs de fréquences normalisées. Pour calculer ces fréquences, j'ai rapporté ces effectifs au nombre total de mots de chaque période. Cela permet de ne pas interpréter les différences de taille entre les sous-corpus comme des différences d'usage. Les résultats de cette analyse diachronique sont présentés dans le Tableau 14, ainsi que dans la Figure 4.

Périodes	Fréquence				Proportion	Ratio
	Total <i>femme</i>	<i>femme_humain</i>	<i>femme_épouse</i>	<i>épouse</i>	<i>femme_épouse</i> / <i>femme</i>	<i>femme_épouse</i> / <i>épouse</i>
1980-1989	0,057 %	0,042 %	0,015 %	0,004 %	26,55 %	3,91
1990-1999	0,068 %	0,049 %	0,018 %	0,004 %	26,87 %	4,94
2000-2009	0,071 %	0,054 %	0,017 %	0,005 %	23,58 %	3,45
2010-2019	0,065 %	0,045 %	0,020 %	0,005 %	31,03 %	3,74
2020-2026	0,066 %	0,049 %	0,018 %	0,004 %	26,75 %	4,06

Tableau 14 : Évolution diachronique des usages de *femme* et d'*épouse* dans *Frantext*.

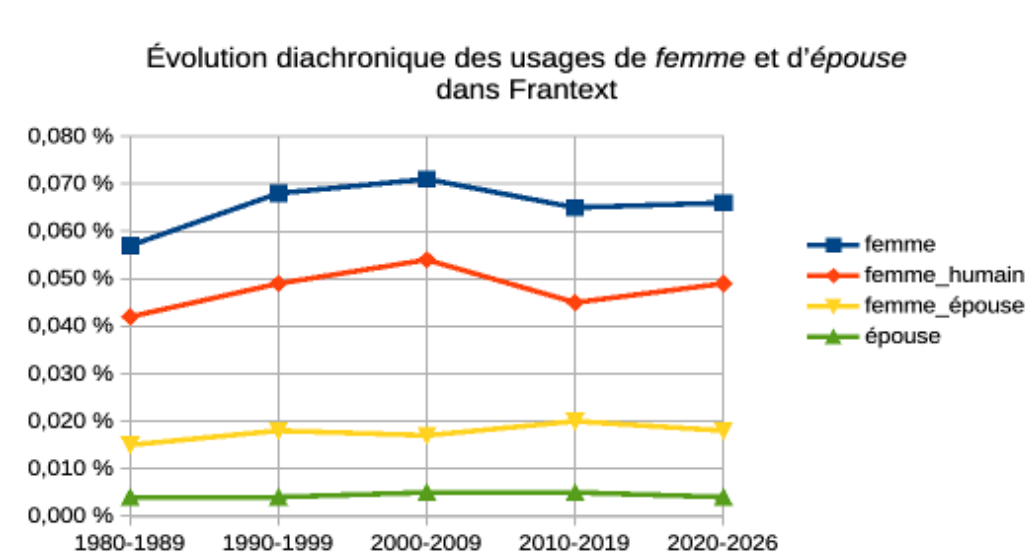


Figure 4 : Évolution diachronique des usages de *femme* et d'*épouse* dans *Frantext*.

Ces représentations permettent d'observer trois types d'informations. D'abord, la fréquence normalisée de *femme* (total de *femme_humain* et *femme_épouse*), de *femme_humain*, de *femme_épouse* et d'*épouse* selon les périodes afin d'observer l'évolution des usages de ces différents termes. Ensuite, la proportion de *femme_épouse* parmi l'ensemble des occurrences de *femme* pour vérifier une éventuelle évolution de l'un des deux sens. Enfin, le rapport entre *femme_épouse* et *épouse*, qui indique combien de fois *femme_épouse* est plus fréquent qu'*épouse*.

Ces résultats permettent de constater plusieurs choses. La fréquence globale de *femme* reste assez stable sur toutes les périodes étudiées, elle varie entre 0,057 % et 0,071 %, sans évolution nette à la hausse ou à la baisse. Le sens *femme_humain* reste toujours bien plus fréquent que *femme_épouse*, quelle que soit la période étudiée. De même, la fréquence de *femme_épouse* reste globalement stable : elle passe de 0,015 % en 1980-1989 à 0,018 % en 2020-2026, avec un pic à 0,020 % en 2010-2019. Quant au terme *épouse*, sa fréquence est très stable et beaucoup plus faible : elle reste autour de 0,004 % à 0,005 % selon les périodes.

En s'intéressant aux proportions, on peut remarquer qu'il n'y a pas de variation visible : la proportion de *femme_épouse* parmi toutes les occurrences de *femme* reste assez stable, autour de 26-27 %, sauf en 2000-2009 où elle descend à 23,58 %, et en 2010-2019 où elle monte à 31,03 %. Cela montre qu'il n'y a pas d'évolution en faveur de l'un des deux sens.

Ensuite, en observant le rapport entre *femme_épouse* et *épouse*, il est clair qu'il n'y a pas de remplacement d'*épouse* par *femme_épouse*. *Femme_épouse* est entre 3,45 et 4,94 fois plus fréquent qu'*épouse* selon les périodes. Dans les années 1980-1989, le ratio est de 3,91 et en 2020-2026, il est de 4,06, ce qui montre une grande stabilité. Toutefois, la période 2020-2026 couvre moins d'années et repose sur un volume de données plus réduit que les autres périodes.

Ces résultats montrent que, dans le corpus contemporain, constitué de textes écrits relevant à la fois de la fiction et de la non-fiction, le terme *femme* conserve une forte propension à désigner l'*épouse*, tandis que *épouse* reste un terme plus spécialisé et moins fréquent. L'hypothèse de départ selon laquelle *femme_épouse* serait de plus en plus utilisé par rapport à *épouse* pour désigner la conjointe n'est pas validée. Il serait cependant intéressant de faire une étude similaire sur des genres de corpus différents, tels que des corpus oraux ou des corpus centrés sur la vie quotidienne, afin de vérifier si cette stabilité se retrouve dans des usages plus spontanés. Il serait également nécessaire d'étendre l'analyse à une période diachronique plus longue, afin de déterminer si des évolutions apparaissent sur le long terme.

Conclusion et perspectives

Bilan

Ce mémoire avait un double objectif : d'une part, traiter la polysémie du mot *femme* en distinguant son sens générique (*femme_humain*) et son sens relationnel (*femme_épouse*) ; d'autre part, observer les usages de ces deux sens dans différents corpus et selon différentes périodes.

Dans l'état de l'art, j'ai d'abord présenté les concepts et les définitions liés à la polysémie, puis la manière dont le TAL appréhende ce phénomène à travers les travaux de Navigli, Bevilacqua, Haber et Poesio. J'ai particulièrement abordé la tâche de WSD, qui permet justement de traiter la polysémie en TAL. Cette partie a montré que la WSD reste une tâche d'actualité, y compris à l'ère des LLM.

Dans la seconde partie, j'ai relevé des régularités d'emploi du mot *femme* dans ses deux sens, afin de constituer une liste exploitable pour créer un système de désambiguïsation. Pour cela, je me suis appuyée sur les dictionnaires, sur l'étude du concordancier dans *Frantext*, ainsi que sur la comparaison des usages de *femme* et d'*épouse* avec *Word Sketch Difference*. J'ai également cherché à voir si le passage au pluriel, avec *femmes*, faisait apparaître des patrons morphosyntaxiques différents, mais cela n'a pas été le cas pour le sens *femme_épouse*.

Dans la troisième partie, j'ai expliqué la méthode utilisée pour créer un système automatique répondant à la problématique initiale, à l'aide d'un programme Python fondé sur les règles linguistiques identifiées et sur la bibliothèque SpaCy. Pour les derniers cas limites, j'ai fait annoter 10 occurrences de *femme* par trois locuteurs français natifs. Cette étape a montré que l'annotation reste une tâche difficile, en particulier lorsque le contexte est limité, comme dans mon système de désambiguïsation, qui se fonde principalement sur le contexte immédiat.

Dans la quatrième partie, l'analyse des traductions a d'abord montré que l'anglais et le japonais sont deux langues intéressantes pour désambiguïser *femme*. Cependant, l'observation du corpus anglais-français *OPUS* a également mis en évidence certaines limites, notamment le fait que de nombreux termes anglophones peuvent traduire *femme*, ce qui nécessite un prétraitement du corpus. J'ai ensuite évalué le programme sur un *gold standard* de 292 occurrences de *femme* à partir du corpus *OPUS*. Le système a obtenu 100 % de précision, 96,58 % de rappel et 98,26 % de F-mesure, ce qui montre la robustesse des règles de désambiguïsation mises en place. Ces scores ont ensuite été comparés à ceux obtenus par un LLM, qui atteint 98,62 % de précision, 99,31 % de rappel et 98,96 % de F-mesure. Cette comparaison montre que les LLM récents peuvent atteindre des performances comparables à celles de systèmes spécialisés de WSD.

Dans la cinquième partie, j'ai appliqué le système à six corpus : *Frantext*, *Wikipédia*, *Le Monde*, *CoMeRe*, *Revues.org* et *Orfeo*, afin de comparer la proportion des deux sens de *femme* selon les genres discursifs et les modalités écrite et orale. J'ai d'abord procédé à une évaluation sur un échantillon de 100 occurrences pour chacun de ces corpus, afin de vérifier que le programme restait performant sur des données différentes. Les résultats montrent une stabilité globale du système, avec des F-mesures comprises entre 97,43 % et 98,99 %.

En ce qui concerne la proportion entre les deux sens de *femme*, le sens *femme_humain* reste dominant dans tous les corpus, avec des scores allant de 72,8 % dans *Frantext* à 97,83 % dans *CoMeRe*. L'analyse

n'a pas non plus montré de différence nette entre l'oral et l'écrit, puisque *Orfeo* reste proche de plusieurs autres corpus avec 79,23 % d'occurrences au sens *femme_humain*. Enfin, l'étude de la fréquence d'apparition des différentes règles de désambiguïsation a montré que les deux règles dominantes sont le classement par défaut au sens *femme_humain* (entre 69,67 % et 92,23 %) et la règle fondée sur les possessifs pour le sens *femme_épouse* (entre 1,55 % et 24,07 %).

Enfin, je me suis intéressée à l'évolution diachronique des usages de *femme* dans le corpus contemporain de *Frantext*. Les résultats montrent une stabilité globale des usages : la fréquence de *femme*, de *femme_humain*, de *femme_épouse* et d'*épouse* varie peu selon les périodes, et le sens *femme_humain* reste toujours largement majoritaire. De même, aucune évolution nette ne permet de conclure à un remplacement progressif d'*épouse* par *femme_épouse*, même si *femme_épouse* demeure systématiquement plus fréquent qu'*épouse*.

En définitive, les résultats obtenus confirment la pertinence du choix de Tanguy et al. (2024, 2025) d'exclure *femme* de leur analyse du lexique de la parenté lorsque le terme n'est pas désambiguïté. Dans l'ensemble des corpus étudiés, le sens *femme_épouse* reste largement minoritaire, jusqu'à ne concerner que 2,17 % des occurrences dans *CoMeRe*. Le système proposé dans ce mémoire permet donc de dépasser cette limite, en isolant automatiquement les occurrences relevant du sens *femme_épouse* et en rendant possible leur réintégration dans une analyse distributionnelle du lexique de la parenté. Pour cela, il faudrait appliquer le système de désambiguïsation au corpus avant l'analyse distributionnelle, puis de ne conserver que les occurrences de *femme* identifiées comme relevant du sens *femme_épouse*. Ces occurrences peuvent ensuite être réintégrées aux côtés des autres termes de parenté dans le corpus d'analyse.

Limites

Ce travail présente toutefois plusieurs limites, qui ont pu influencer les résultats. Tout d'abord, l'évaluation des modèles repose généralement sur des annotations humaines, qui peuvent elles-mêmes comporter une part d'incertitude. Dans mon cas, certains choix d'annotation se sont révélés discutables a posteriori : il m'est arrivé de classer une occurrence dans un sens, avant de constater par la suite qu'un autre classement aurait pu être envisagé. Par exemple, j'ai systématiquement classé *femme au foyer* dans le sens *femme_humain*, alors que cette expression peut également relever du sens *femme_épouse* selon le contexte.

Ainsi, il aurait été préférable de faire annoter par plusieurs personnes l'ensemble des occurrences utilisées pour l'évaluation des performances du programme sur les différents corpus, plutôt que de m'appuyer uniquement sur ma propre annotation. Cependant, cela aurait représenté plus de 1200 occurrences à annoter et donc un travail important pour plusieurs annotateurs. Pour cette raison, il n'a pas été possible de mettre en place une telle évaluation dans le cadre de ce mémoire.

Par rapport au programme lui-même, il aurait été intéressant d'utiliser une ressource externe ou une API lexicale afin d'enrichir automatiquement la liste des métiers et des rôles sociaux. J'aurais également aimé trouver une solution pour traiter les occurrences de type *avoir une femme*, que j'ai finalement choisi d'exclure du programme, car elles entraînaient trop d'ambiguïtés selon les contextes. Une piste aurait été d'utiliser la présence ou l'absence de *y* pour distinguer des phrases comme *il y avait une femme* et *il a une femme*. De même, j'ai tenté de mettre en place un système de pondération à partir de mes connaissances en informatique, afin de compléter les règles linguistiques mises en place, mais cette piste n'a pas abouti en raison de mes compétences techniques actuelles.

Enfin, l'analyse du corpus *OPUS* a montré que l'anglais comporte de nombreux termes pouvant correspondre à *femme*. Il aurait donc été possible d'envisager une autre langue dans le cadre d'un corpus parallèle, afin de limiter la diversité des équivalents possibles et d'obtenir davantage d'occurrences exploitables. La langue japonaise constituait également une piste intéressante, que j'aurais aimé explorer plus en détail dans ce travail.

Perspectives

Ce travail pourrait être prolongé de plusieurs manières. D'abord, concernant l'évolution diachronique, il serait pertinent de poursuivre l'analyse sur des corpus relevant de genres différents, notamment des corpus oraux ou centrés sur la vie quotidienne, plutôt que sur des corpus littéraires, afin de vérifier si les tendances observées se confirment.

J'aurais également aimé m'intéresser plus finement aux usages de *femme_humain* et de *femme_épouse* en observant leurs cooccurrents, afin de voir comment ces deux sens sont employés selon les genres de corpus. Il serait par exemple intéressant d'extraire les cooccurrents associés à chaque sens sous forme de tables de fréquence, puis de les analyser pour dégager les thématiques propres à chaque genre discursif.

Par ailleurs, nous avons vu dans la section 2.3, avec *Word Sketch Difference*, que les termes *femme* et *épouse* ne semblent pas être utilisés dans le même niveau de formalité : *épouse* paraît associé à des contextes plus formels. Nous pourrions analyser plus en détail les occurrences de *femme_épouse* et d'*épouse*, afin de vérifier si cette différence se confirme. De plus, nous pourrions vérifier si les usages respectent les prescriptions selon lesquelles *femme_épouse* serait plus délicat pour désigner sa conjointe que *épouse* (Pruvost, 2024). Enfin, nous pourrions comparer les usages de *femme_épouse* et d'*épouse* avec ceux de *mari* et d'*époux* afin de voir quels termes se rapprochent le plus.

Par rapport aux autres termes de la parenté, les observations et analyses réalisées dans ce mémoire pourraient servir de base méthodologique pour la désambiguïsation d'autres termes polysémiques, comme *filles*. Nous pourrions notamment tester l'idée d'un système plus simple, fondé sur les possessifs pour repérer les emplois parentaux, sur des exceptions lexicales comme *filles au pair*, puis sur un classement par défaut des autres occurrences dans le sens d'*enfant de sexe féminin*. Comme je l'ai déjà suggéré dans la section 5.1.5, une telle approche pourrait permettre de limiter les biais dans l'analyse distributionnelle, sans passer par une longue étude de corpus ni par le recours à un LLM.

Bibliographie

- Apidianaki, M. (2006). *Traitement de la polysémie lexicale dans un but de traduction*. Actes de TALN 2006, Leuven, Belgique.
- Apidianaki, M. (2023). From word types to tokens and back: A survey of approaches to word meaning representation and interpretation. *Computational Linguistics*, 49(2), 465-523.
- Baider, F, et Jacquy, E. (2008). Modélisation sémantique, sens différentiel et genre, *Congrès mondial de Linguistique française* (EDP Sciences), Paris, France, pp. 177.
- Benhamou, S. (1986). Analyse dictionnaire de « femme » et de « homme », *Cahiers de lexicologie*, (48), 27-67.
- Benzitoun, C., & Cappeau, P. (2025). Les corpus et leur exploitation. *Encyclopédie Grammaticale du Français*, en ligne : encyclogram.fr.
- Bertin, Thomas. (2018). *La polysémie des noms de parties du corps humain en français : analyse sémantique de artère, bouche, cœur, épaule et pied*. Thèse de doctorat, Université de Normandie.
- Bevilacqua, M., Pasini, T., Raganato, A., et Navigli, R. (2021). Recent Trends in Word Sense Disambiguation: A Survey. *Thirtieth International Joint Conference on Artificial Intelligence (IJCAI 2021)*, Montréal, Canada, pp. 4330-4338.
- Bréal, M. (1897). *Essai de sémantique (science des significations)*. Paris, France : Hachette.
- Caldarola, E. G., Picariello, A., et Rinaldi, A. M. (2016). WordNet exploration and visualization in Neo4J. A tag cloud based approach. *The Eighth International Conference on Information, Process, and Knowledge Management (eKNOW 2016)*, Venice, Italy.
- Forgar, S. (2020, 10 novembre). *En 2020, le dictionnaire Oxford change enfin sa définition datée et sexiste du mot « femme »*. Madame Figaro.
- Fuchs, C. (1996). *Les ambiguïtés du français*. Ophrys.
- Gold, D. L. (1986). Ordering the senses in a monolingual dictionary entry. *Babel*, 32(1), p. 44-49.
- Gries, S. T. (2019). Polysemy. In: Dąbrowska, E., et Divjak, D. (eds.), *Cognitive Linguistics - Key Topics*, De Gruyter Mouton, p. 23-43.
- Habert, B. (2005). *Instruments et ressources électroniques pour le français*. Ophrys.
- Haber, J., et Poesio, M. (2024). Polysemy - evidence from linguistics, behavioral science, and contextualized language models, *Computational Linguistics*, 50(1), 351-417.

- Jakubiček, M., Kilgarriff, A., Kovář, V., Rychlý, P., & Suchomel, V. (2013). *The TenTen corpus family*. In *Proceedings of the 7th International Corpus Linguistics Conference (CL 2013)*. Lancaster, UK, pp. 125-127.
- Kilgarriff, A. (1997). I Don't Believe in Word Senses. *Computers and the Humanities*, 31(2), 91–113.
- Kilgarriff, A., Rychlý, P., Smrž, P., et Tugwell, D. (2004). *The Sketch Engine*. Actes du 11e Congrès international EURALEX (EURALEX 2004), Lorient, France, pp. 105-116.
- Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. Chicago : University of Chicago Press.
- Le Draoulec, A., & Péry-Woodley, M.-P. (2026, 4 juin). *D'une femme l'autre*. Bling. Consulté le 4 juin 2026, à l'adresse <https://bling.hypotheses.org/10609>
- Le Draoulec, A., & Rebeyrolle, J. (2025). Fille de, femme de. Dans A. Le Draoulec & M.-P. Péry-Woodley (dir.), *Entre elle et lui : Variations sur les asymétries de genre en français* (pp. 173-178). Éditions de l'Aube.
- Lison, P., & Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 923-929).
- Loureiro, D., Rezaee, K., Pilehvar, M. T., & Camacho-Collados, J. (2021). *Analysis and Evaluation of Language Models for Word Sense Disambiguation*. arXiv:2008.11608.
- Luan, Y., Hauer, B., Mou, L., & Kondrak, G. (2020a). Improving Word Sense Disambiguation with Translations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 4055-4065).
- Michel Mathieu-Colas. (1995). *Représentation de la polysémie dans un dictionnaire électronique. Lexicomatique et dictionnaires*, Lyon, France, pp. 317–325.
- Mazaleyrat, H. (2010). *Vers une approche linguistico-cognitive de la polysémie : Représentation de la signification et construction du sens*. Thèse de doctorat, Université Blaise Pascal - Clermont-Ferrand II.
- Mikolov, T., Chen, K., Corrado, G., et Dean, J. (2013). *Efficient estimation of word representations in vector space*. arXiv preprint arXiv:1301.3781.
- Miller, G. A., Leacock, C., Teng, R., & Bunker, R. T. (1993). A semantic concordance. In *Proceedings of the workshop on Human Language Technology* (pp. 303-308).
- Morin, E., Dufour-Kowalski, S., & Daille, B. (2004). *Extraction de terminologies bilingues à partir de corpus comparables*. In Actes de TALN (pp. 19–21).
- Navigli, R. (2009). *Word sense disambiguation: A survey*. *ACM Computing Surveys*, 41(2), 1-69.

- Navigli, R. (2026). Is word sense disambiguation dead in the LLM era? *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(46), pp. 39753–39762, Singapour.
- Picoche, J. (1989). *Polysémie n'est pas ambiguïté. Cahiers de praxématique*, 12, 75-89.
- Pruvost, J. (2021). *Les dictionnaires français, outils d'une langue et d'une culture*. Ophrys.
- Pruvost, J. (2024). *Le féminin : au fil des mots et de l'histoire*. Tallandier.
- Raganato, A., Camacho-Collados, J., & Navigli, R. (2017). Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers* (pp. 99-110). Association for Computational Linguistics.
- Salle, M. (2023). Quand le mot « femme » disparaît. Innovations sémantiques ou nouvelles formes d'invisibilisation ? Dans Lemarchand, P., et Salle, M. (dir.) *Qu'est-ce qu'une femme ? Catégories homme/femme : débats contemporains*, Éditions Matériologiques, p. 69-91.
- Siméone, C. (2019, 8 février). De « femelle de l'homme » à « femme fatale », l'évolution du mot « femme » à l'Académie française depuis 1694. *Radio France*.
- Śmigielska, B. (2023). *Traduction automatique et désambiguïsation des sens des mots. Le cas du verbe français louer*. *Neophilologica*, 35, 1-26.
- Tanguy, L., Fabre, C., Hathout, N., et Ho-Dac, L.-M. (2024). Le sens de la famille : analyse du vocabulaire de la parenté par les plongements de mots. *Actes des 17es Journées internationales d'Analyse statistique des Données Textuelles (JADT 2024)*, Bruxelles, Belgique.
- Tanguy, L., Fabre, C., Hathout, N., et Ho-Dac, L.-M. (2025). Embeddings, topic models, LLM : un air de famille. *Actes de la conférence TALN 2025*, Marseille, France, pp. 295-312.
- Victorri, B., et Fuchs, C. (1996). *La polysémie : construction dynamique du sens*. Paris, France : Hermès.
- Weaver, W. (1949). *Translation*. In W. N. Locke et A. D. Booth (Eds.), *Machine translation of languages: Fourteen essays* (pp. 15-23). Cambridge, MA: MIT Press.

Annexes

Annexe A : Liste des locutions de *femme*

Locutions au sens être humain de sexe féminin	Locutions au sens épouse
<i>femme du monde, femme ingénieur, femme enfant, femme de tête, maîtresse femme, femme au foyer, femme fatale, femme-objet, sage-femme, bonne femme, femme mariée, femme divorcée, femme veuve, femme entretenue, femme de ménage, femme de chambre, femme de service, femme de lettres, femme de réconfort, femme enceinte, femme de science, femme politique, femme d'affaires, femme médecin, femme libérée, femme libre, femme mûre, femme célibataire, femme de loi, femme d'État, femme de pouvoir, femme d'esprit, femme de caractère, femme de goût, femme de compagnie, femme de charge, femme légère, femme facile</i>	<i>femme fidèle, femme infidèle, femme adultère, femme trompée, femme bafouée, femme chérie, femme bien-aimée, femme légitime, femme aimante, femme morganatique, femme répudiée, défunte femme, tendre femme, future femme, charmante femme, chaste femme, ex-femme, femme promise</i>

Annexe B : Listes de régularités autour de *femme* et *femmes* au sens humain

Régularités autour de <i>femme_humain</i>	
Contexte gauche	Contexte droit
<i>Homme-femme / homme/femme / homme et femme / homme ou femme / ni homme ni femme Femme-femme Femme à femme Homme à femmes Amie femme Apparence femme Aucune / autre / nulle / toute / seule femme Cette/une/la femme Cette femme Chaque femme Chercher femme Collection femme Corps/main/voix/visage de femme (nom + de + femme) Devenir / être / naître / redevenir / rester femme Femmes entre elles Moitié en femme / transformer / déguiser / habiller en femme Se/me/te conduire en femme Se/me/te sentir femme</i>	<i>Cette femme-là Femme active / admirable / adorée / âgée (femme + [assez] + adjectif) Femme + nationalité Femme + au(x) + nom De femme en femme Femmes et enfants Femmes et hommes</i>

<i>Belle / brave / célèbre / excellente / faible / grande / grosse / honnête / jeune / jolie / pauvre / petite / sacrée / sainte / sale / simple / superbe / vieille / vraie / âgée + femme (adjectif + femme)</i> <i>Genre femme</i> <i>Moins/peu/plus femme</i> <i>Nom + pour + femme</i> <i>En tant que femme + adjectif</i> <i>Quelle/quelque femme</i> <i>Une telle femme</i> <i>Nombreuses / quelques / sans femmes</i> <i>Beaucoup /peu de femmes</i> <i>Égalité femmes</i>	
---	--

Annexe C : Listes de régularités autour de *femme* et *femmes* au sens épouse

Régularités autour de <i>femme_épouse</i>	
Contexte gauche	Contexte droit
<i>Mari-femme / mari et femme</i> <i>Ma / ta / sa / votre / leur + femme</i> <i>Son / mon / votre / ton ancienne / autre + femme</i> <i>Ma / ta / sa / votre / leur + première / deuxième / seconde / troisième / quatrième / nouvelle / propre / future femme</i> <i>Ma / ta / sa / votre / leur + adjectif + femme</i> <i>Nom + dont la femme</i> <i>Avoir femme et/ou enfants</i> <i>Avoir une femme</i> <i>Prendre / choisir pour</i> <i>Prendre / quitter femme</i>	<i>Femme chérie</i> <i>La + femme + de + nom propre</i> <i>Femme + de/d'un + [adjectif] + nom de métier</i> <i>Femme d'expatrié / de prisonnier</i> <i>Femme de mon + nom de métiers</i> <i>La femme + du + nom de métier</i>

Annexe D : Arbre de décision des règles de désambiguïsation

"femme" est présent dans le texte ?

└─ Le token est "ex-femme" ou "mari-femme" ?

└─ Oui → épouse

└─ Le token est "femme" ?

└─ Oui → appliquer les règles suivantes dans l'ordre :

└─ **Locution parmi une liste (*femme + de + nom*) ?**

Exemples : *femme de ménage, femme de chambre...*

└─ Oui → être humain de sexe féminin

└─ **"femme" a-t-il un déterminant possessif comme dépendant ?**

Exemples : *ma femme, sa nouvelle femme, son ex femme...*

└─ Oui → épouse

└─ **Locution parmi une liste (*femme + nom ou adjectif*) ?**

Exemples : *femme politique, femme mariée...*

└─ Oui → être humain de sexe féminin

└─ **Séquence "femme au foyer" ?**

└─ Oui → être humain de sexe féminin

└─ **Le mot précédent est "maîtresse" ?**

└─ Oui → être humain de sexe féminin

└─ **Quantifieur avant "femme" ?**

Exemples : *certaines femmes, quelques femmes, peu de femmes, beaucoup de femmes...*

└─ Oui → être humain de sexe féminin

└─ **Verbe "prendre" ou "quitter" avant "femme" ?**

└─ Oui → épouse

└─ **Construction avec *verbe + préposition + femme* ?**

Exemples : *prendre pour femme, choisir comme femme...*

└─ Oui → épouse

└─ **Mot après "femme" parmi une liste ?**

Exemples : *femme légitime, femme adultère, femme infidèle, femme répudiée...*

└─ Oui → épouse

└─ **Mot avant "femme" = "future" ou "ex" ?**

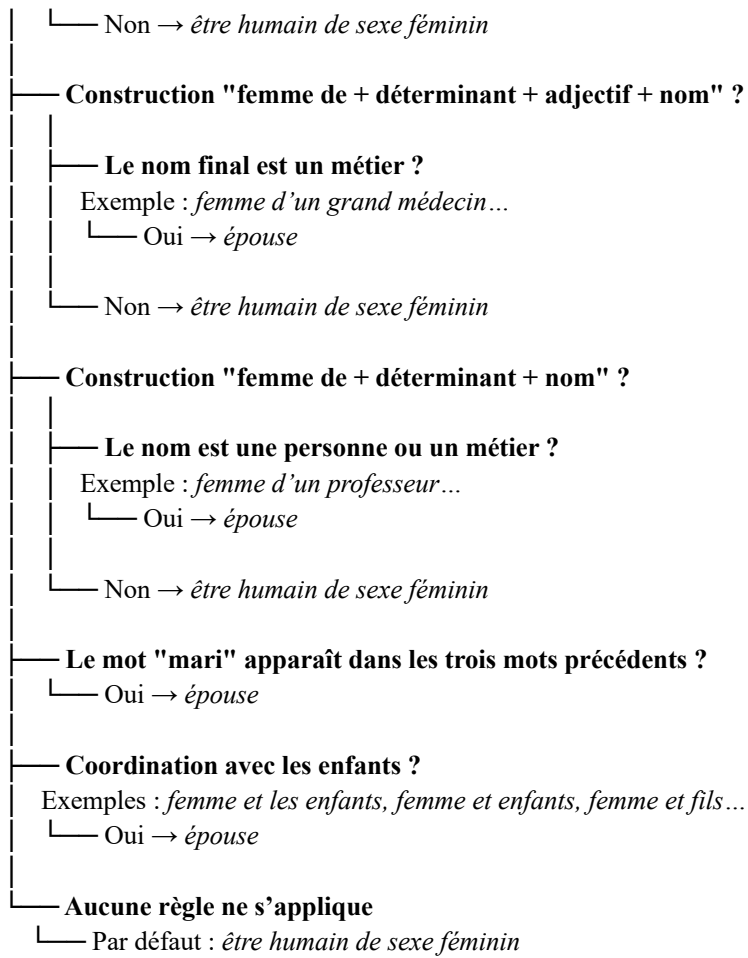
└─ Oui → épouse

└─ **Construction "femme de + nom / nom propre" ?**

└─ **Le nom après "de" est une personne, un métier, "monsieur" ou "madame" ?**

Exemples : *femme de médecin, femme de monsieur X...*

└─ Oui → épouse



Annexe E : Guide d'annotation du mot *femme*

Dans le cadre de mon mémoire en Traitement Automatique des Langues, je travaille sur la polysémie du mot *femme*, cela désigne le fait qu'un même mot peut avoir plusieurs sens liés entre eux, dont l'interprétation dépend du contexte.

Femme a deux sens principaux qui sont *être humain de sexe féminin* et *épouse*. Pour rappel, voici les définitions de ces deux sens de *femme* issues du dictionnaire *Le Robert en ligne*.

Être humain de sexe féminin : adulte de sexe féminin.

Épouse : femme unie par le mariage.

Je cherche à distinguer les deux sens dans différentes phrases afin de recueillir plusieurs points de vue et m'aider à décider quelles règles linguistiques peuvent m'aider en vue de créer un programme automatique de désambiguïsation du mot *femme*.

J'ai besoin de ton aide pour annoter quelques phrases contenant le mot *femme*. Tu dois t'aider du contexte pour déterminer si tu penses qu'il s'agit du sens *être humain de sexe féminin* ou *épouse*. Dans commentaire, tu dois expliquer pourquoi tu as choisi un sens plutôt que l'autre. Par exemple, tu peux dire qu'un mot de la phrase t'a aidé. **En dernier recours**, tu peux également dire que pour toi les deux sens peuvent convenir en justifiant ce choix.

Il n'y a pas de mauvaises réponses, mon objectif est justement de recueillir plusieurs avis afin de déterminer si les indices linguistiques présents dans ces phrases suffisent à déterminer un sens ou l'autre. Il n'y a pas de limite de temps.

Voici un exemple d'annotation :

Phrase : *C'est une femme enceinte.*

Sens : *être humain de sexe féminin.*

Commentaire : « *une* » montre qu'on parle d'une femme en général, pas de la femme de quelqu'un.

Voici les phrases à annoter :

Phrases	Humain	Épouse
C'est une femme fidèle.		
La défunte femme aurait senti la mort arriver.		
C'est une femme bafouée.		
La femme et les enfants sont partis le week-end à Paris.		
La femme et les gosses.		
C'est une chaste femme.		
C'est une femme infidèle.		
Oui, je devinai ce que pouvait ressentir une femme adultère.		

Annexe F : Nombre d'occurrences de *femme* et *épouse* dans *Frantext* selon cinq périodes

Périodes	Nombre de mots	Nombre d'occurrences			
		Total <i>femme</i>	<i>femme_humain</i>	<i>femme_épouse</i>	<i>épouse</i>
1980-1989	14 017 203	7976	5850	2126	544
1990-1999	10 459 449	7070	5170	1900	385
2000-2009	11 990 728	8490	6488	2002	580
2010-2019	9 504 532	6206	4280	1926	515
2020-2026	3 726 964	2471	1810	661	163
Total	49 698 876	32 213	23 598	8615	2187

Déclaration sur l'honneur de non-plagiat

(à joindre au mémoire à la fin du document)

Je soussigné.e,

Nom, Prénom : **PIERRON MATHILDE**

Régulièrement inscrit.e à l'Université de Toulouse II Jean Jaurès

N° étudiant : **22001920**

Année universitaire : **2025 - 2026**

certifie que le document joint à la présente déclaration est un travail original, que je n'ai ni recopié ni utilisé des idées ou des formulations tirées d'un ouvrage, article ou mémoire, en version imprimée ou électronique, sans mentionner précisément leur origine et que les citations intégrales sont signalées entre guillemets.

Fait à : **Toulouse**

Le : **12/06/2026**

Signature :

