Construction et analyse d'un corpus de science-fiction pour l'étude des mots-fictions

Mathilde Ducos et Frédéric Landragin

Abstract. La science-fiction est un genre littéraire qui se distingue des autres par sa faculté à concevoir un nouveau lexique lors de l'écriture des récits, constitué de termes appelés mots-fictions. Cependant, le nombre de corpus de récits de science-fiction est à ce jour très limité, et aucun d'eux ne permet un accès aux textes intégraux. De plus, les chercheurs en linguistique et en littérature extraient à la main les mots-fictions, rendant les analyses de ces termes plus longues et fastidieuses. C'est pourquoi nous proposons NovSFcorpus[1], un corpus permettant d'accéder à environ 250 textes intégraux de romans et de nouvelles de science-fiction francophones libres de droits, contenant 10 millions de tokens et annotés en mots-fictions grâce à une méthodologie semi-automatique. Plusieurs catégories de mots-fictions ont alors été mises en avant pour une compréhension plus fine de leur construction. Un modèle de reconnaissance de mots-fictions pourra ensuite être entraîné sur ce corpus afin de détecter automatiquement ces termes étudiés par plusieurs disciplines de la linguistique.

1 Introduction

La science-fiction francophone constitue un champ d'étude particulièrement riche pour l'analyse linguistique et narratologique. Définie comme un ensemble de récits spéculatifs par Simon Bréan [1] explorant des hypothèses scientifiques ou une anticipation technologique, elle repose sur le principe central de la construction d'un monde où un élément modifié permet d'énoncer la question « et si ? ». Et si cela était possible ? Et si les choses n'étaient pas ce qu'elles semblent être ? Et si l'on pensait autrement ? [2]. Cette littérature mobilise des mécanismes variés, allant de la suspension consentie d'incrédulité du lecteur (selon laquelle le lecteur consent mettre de côté ses connaissances du monde le temps de la lecture, concept inventé par Samuel Coleridge en 1817 dans *Biographia Literia*) à l'encyclopédie spécifique à la science-fiction (l'ensemble du lexique du lecteur au contact d'un récit [3]).

Un concept clé pour décrire les mécanismes fictionnels propres à la science-fiction est celui de novum [4], ou mot-fiction, traduction du terme [5] : un élément lexical désignant une entité, un objet ou un phénomène inexistant dans le monde réel, mais cohérent dans l'univers du récit. Un exemple classique est « télécran » dans *1984* de George Orwell (1949). Ces unités jouent un rôle structurant, car elles condensent la différence ontologique entre le monde de la fiction et celui de référence, c'est-à-dire celui du lecteur, tout en participant à la construction de la cohérence interne du récit.

[1] Corpus : https://github.com/mducos/FrenchSFcorpus/tree/main/data/NovSFcorpus
Codes réutilisables : https://github.com/mducos/FrenchSFcorpus/tree/main/script/nov_annotations

L'étude de tels expressions dans un corpus francophone se justifie à plusieurs niveaux. Sur le plan littéraire, la science-fiction française possède une histoire spécifique, distincte du canon anglo-américain par les thèmes étudiés [1], et possède de multiples appellations : « roman scientifique », « roman d'hypothèse », « merveilleux scientifique », « anticipation » [6]. Ce genre a suivi sa propre trajectoire, avec des évolutions et des esthétiques spécifiques. Sur le plan linguistique, les récits francophones de science-fiction constituent un terrain riche pour étudier la créativité lexicale, les stratégies de création de mots-fictions, ainsi que leur intégration dans la syntaxe et le discours. De plus, les corpus actuels de science-fiction française ont une disponibilité limitée et ne s'étendent pas à une étude des mots-fictions, limitant le développement d'outils adaptés à leur analyse.

Dans ce contexte, cet article présente un corpus original de récits francophones de science-fiction, constitué de textes libres de droits publiés entre 1860 et 1940 (période considérée comme l'âge d'or de la science-fiction française [7]). Ce corpus est enrichi d'annotations portant sur les mots-fictions, réalisées selon une méthodologie semi-automatique et destinées à supporter diverses recherches sur ce phénomène propre au genre. Les textes intégraux et les annotations sont mis à disposition en accès libre afin de favoriser la reproductibilité des analyses et de soutenir le développement d'outils linguistiques appliqués à la littérature.

Les contributions de cet article sont doubles :

- d'une part, développer une méthodologie pour la caractérisation des mots-fictions ;
- d'autre part, documenter la méthodologie de la construction et de l'annotation du corpus.

Nous détaillons successivement l'état de l'art (Section 2), la méthodologie d'acquisition et d'annotation du corpus (Section 3), puis les caractéristiques quantitatives et qualitatives du corpus, ainsi que des analyses exploratoires (Section 4).

2 Etat de l'art

2.1 Définition de la science-fiction

L'appellation de science-fiction soulève des difficultés théoriques et historiques qui empêchent toute définition immédiate fondée sur les seules pratiques éditoriales. Une première solution consisterait à s'appuyer sur les collections d'éditeurs explicitement dédiées à la science-fiction. Cette approche se heurte cependant à un obstacle majeur : de nombreux romans considérés aujourd'hui comme des œuvres centrales du genre, comme *1984* de George Orwell (1948) ou les récits de Jules Verne, sont publiés dans des collections généralistes ou patrimoniales (*1984* est par exemple publié dans la collection poche généraliste des éditions Gallimard, Folio, et non dans leur collection spécialisée en science-fiction : Folio SF). Les frontières éditoriales ne coïncident donc pas avec les frontières du genre, ce qui limite la pertinence d'un critère strictement institutionnel pour en fournir une définition.

Une seconde option consiste à définir la science-fiction selon des critères littéraires et linguistiques. Dans cette perspective, plusieurs théories situent le cœur du genre dans la présence de mots-fictions qui déclenchent une distanciation cognitive (ou *cognitive estrangement* en anglais) ([4], [8]). Cet écart cognitif distingue les genres de l'imaginaire constitués de la fantasy et de la science-fiction des autres formes narratives, en articulant simultanément une distanciation rationnelle et une transformation du monde représenté. La science-fiction se singularise alors par la nature des mondes qu'elle met en scène : selon Simon Bréan, les récits reposant sur un « régime spéculatif » postulent que le monde fictionnel pourrait se substituer au monde de référence, tandis que la fantasy relève d'un «

régime extraordinaire » fondé sur une rupture explicite entre ces deux mondes. Une autre manière de formuler cette distinction consiste à considérer que les mots-fictions de la science-fiction s'ancrent dans des moyens scientifiques ou technologiques, tandis que ceux de la fantasy reposent sur la magie [4]. Les deux approches convergent pour identifier la science-fiction comme un espace de spéculation rationnelle fondé sur la transformation plausible du réel.

La notion d'anticipation apparaît ensuite comme un sous-genre de science-fiction. Les travaux récents, notamment dans le cadre du projet ANR Anticipation[2], montrent que ce terme recouvre un ensemble hétérogène qui excède l'acception strictement contemporaine de la science-fiction. L'émergence du roman d'anticipation s'inscrit dans un contexte culturel précis : dès les années 1860, la consolidation d'une culture médiatique de la vulgarisation, illustrée par la collaboration entre Jules Verne et Pierre-Jules Hetzel dans le Magasin d'éducation et de récréation, fournit un cadre propice à la circulation des savoirs et à l'articulation entre fiction et discours scientifiques. À partir des années 1880, l'essor spectaculaire de la presse quotidienne et la multiplication des collections de littérature populaire favorisent la diffusion massive de récits spéculatifs qui participent à la mise en forme des imaginaires scientifiques de la fin du XIXe et du début du XXe siècle [9].

L'étude du roman d'anticipation sur une période allant des années 1860 à 1940 permet de saisir l'évolution d'un corpus caractérisé par sa diversité formelle et conceptuelle. La Seconde Guerre mondiale marque à la fois un renouveau des imaginaires scientifiques et une transformation profonde des supports éditoriaux, ce qui justifie la clôture du champ à cette date. Alors que la littérature de science-fiction bénéficie d'une tradition de recherche solide dans le monde anglophone (notamment grâce à des revues académiques spécialisées telles que *Science Fiction Studies* ou *Foundation – The International Review of Science Fiction*), elle demeure encore peu explorée dans l'espace francophone.

Pourtant, la science-fiction littéraire est riche d'innovations narratologiques, stylistiques, et même lexicales avec la création de néologismes particuliers qui sont au cœur de cet article.

2.2 Définition des mots-fictions

Télécran (*1984*, George Orwell, 1949), *psychohistoire* (*Cycle de Fondation*, Isaac Asimov, 1942-1993), *terraformation* (*Trilogie de Mars*, Stanley Robinson, 1992-1996). Ces mots sont inconnus et pourtant nous comprenons aisément leur signification car ils sont construits selon une « logique cognitive » [4]. Certains sont même entrés dans le langage courant, comme *robot*[3], terme utilisé pour la première fois par Karel Čapek dans *R. U. R.* en 1920. Ainsi, grâce à nos connaissances de la langue, nous pouvons déduire que *télécran* décrit un objet de diffusion d'images à distance sur un écran, *psychohistoire* est l'étude de l'histoire à travers la psychologie humaine, et *terraformation* est le processus de transformation d'un milieu naturel afin de le rendre semblable à celui de la Terre.

Cette capacité d'interprétation des mots nouveaux s'inscrit plus largement dans les travaux linguistiques consacrés à la néologie. Les néologismes sont généralement définis comme des unités lexicales nouvelles ou comme des sens nouveaux attribués à des formes existantes [10], mais cette définition apparente masque la complexité des mécanismes en jeu. Comme le montrent les analyses en lexicologie de Semprun [11], la création néologique répond à des besoins variés : nommer des concepts nouveaux, nommer des concepts préexistants mais n'ayant jamais été nommés, ou encore nommer des concepts

[2] Site officiel du projet : https://anranticip.hypotheses.org/
[3] Académie française. "Robot" *Dictionnaire de l'Académie française*, 9e éd., https://www.dictionnaire-academie.fr/article/A9R2773. Consulté 26 mars 2026.

plus anciens avec un œil moderne. Les néologismes sont un procédé courant en littérature : Boris Vian utilise des mots-valises dans *L'Ecume des jours* (1947) avec *députodrome* ou *bedon*, ou encore dans *L'Herbe rouge* (1950) avec *sarcastifleur* ou *tertreux*. Mais plus particulièrement en science-fiction, ce processus prend une dimension spécifique : les mots-fictions ne se contentent pas de désigner, ils participent à la construction d'univers cohérents et plausibles, tout en mobilisant des schémas morphologiques et sémantiques familiers qui facilitent leur compréhension par le lecteur.

Ces néologismes, bien qu'inexistants dans un dictionnaire officiel et ne possédant donc pas de définition, sont compréhensibles grâce au partage d'un vocabulaire commun entre auteur et lecteur. Cependant, ce qui distingue la science-fiction d'autres genres littéraires est la généralisation de l'utilisation de ces néologismes. En effet, « la science-fiction se distingue par la prédominance narrative ou l'hégémonie d'un « novum » fictif (nouveauté, innovation) validée par la logique »[4] [4]. Le novum est donc une nouveauté diégétique appartenant au monde de fiction, sans référent dans le monde de référence. Il est par ailleurs central en science-fiction car il « implique un changement de l'univers entier du récit, ou du moins de certains aspects cruciaux de celui-ci »[5] [4].

En 1982, Marc Angenot traduit le mot novum en mots-fictions et précise sa définition en affirmant que ces créations lexicales de la science-fiction désignent soit « 1) les mots censés anticiper sur un état futur de la langue du récit ou censés relever d'un univers linguistique parallèle, 2) les mots qu'on suppose empruntés à une langue extraterrestre » [5]. Autrement dit, les mots-fictions sont des signifiants nouveaux ayant des référents imaginaires.

En les intégrant dans le récit, ces mots-fictions se lexicalisent et permettront au lecteur d'étoffer son encyclopédie, concept introduit par Umberto Eco en 1979 pour désigner l'ensemble du lexique du lecteur au contact d'un récit de fiction [3]. En particulier pour la science-fiction, Richard Saint-Gelais précise ce terme en 1999 en le nommant xénoencyclopédie à cause des mots-fictions désignant des référents imaginaires [12].

Plusieurs thèses de doctorat ont été publiées sur une étude de ces mots-fictions, dans différents champs disciplinaires comme la lexicologie, la traductologie ou encore la sociologie. [13] adopte une perspective pluridisciplinaire pour examiner la traduction des mots-fictions : en mobilisant des outils issus de la linguistique, de la traductologie et de la sociologie, elle analyse les mécanismes qui régissent le passage des termes inventés de l'anglais au français et met en évidence les transformations objectives opérées lors de ce transfert. La thèse de Gindre [14], ancrée en lexicologie, s'intéresse à la formation des mots-fictions dans un corpus de science-fiction anglophone. Elle montre que, malgré une répartition des matrices lexicogéniques comparable à celle du lexique réel, les œuvres de science-fiction se distinguent par des combinaisons inédites, des réseaux sémantiques complexes et la nécessité d'une double lecture ancrée à la fois dans l'univers fictionnel et dans les attentes du lectorat. Cette approche empirico-inductive met en lumière la cohérence interne des lexies de science-fiction et souligne l'intérêt d'analyses systématiques pour comprendre les dynamiques lexicales propres au genre. Cependant, pour ces deux études, les analyses ont porté sur un ensemble de mots-fictions récupéré manuellement dans les romans de science-fiction les plus connus du grand public durant une certaine période, et traité en dehors de leur contexte narratif.

[4] Notre traduction de « SF is distinguished by the narrative dominance or hegemony of a fictional « novum » (novelty, innovation) validated by cognitive logic »
[5] Notre traduction de « entails a change of the whole universe of the tale, or at least of crucially important aspects thereof »

2.3 Les corpus spécialisés en littérature et en science-fiction

Plusieurs corpus peuvent servir de point d'appui à l'étude des récits littéraires francophones de science-fiction. Toutefois, aucun ne propose des ressources répondant simultanément aux contraintes de disponibilité des textes, de distinction des genres littéraires entre la science-fiction et les autres et d'annotation des mots-fictions. Cette section présente les principales ressources existantes et évalue leur pertinence dans ce cadre.

Le corpus Chapitres issu du projet ANR éponyme[6] propose un large ensemble de romans couvrant la période 1789–1939, incluant des œuvres canoniques et non canoniques. Bien que certains récits appartiennent à la science-fiction, leur identification est rendue difficile par l'hétérogénéité des métadonnées : certaines œuvres explicitement rattachables à la science-fiction ne sont pas étiquetées comme telles, tandis que d'autres le sont. Ces petites erreurs ont des conséquences sur les exploitations de traitement automatique des langues et complique l'extraction d'un sous-corpus spécialisé. Par ailleurs, seuls vingt-cinq titres relevants de la science-fiction y sont recensés, ce qui constitue un volume insuffisant pour former un corpus utilisable pour de la linguistique de corpus et du traitement automatique des langues.

La bibliothèque du projet Gutenberg[7] constitue également une source de textes libres, organisée en grandes catégories thématiques et linguistiques. Les romans francophones y sont dispersés entre plusieurs rubriques, notamment « SF-Fantasy » et « French », ce qui impose un travail de filtrage manuel important pour isoler les récits pertinents. De plus, la couverture générique et temporelle demeure limitée pour la période d'étude considérée : de nombreux récits parus en revue ou des œuvres aujourd'hui moins connues ne sont pas présents dans la base. Ces contraintes rendent l'extraction d'un corpus spécialisé difficile et chronophage.

Les corpus PhraseoRomSF-fr et PhraseoAnticipation-fr [15], développés dans le cadre du projet ANR PhraseoRom, représentent des ressources spécifiquement orientées vers la science-fiction et l'anticipation. PhraséoRomSF-fr rassemble des romans français publiés entre 1952 et 2014, pour un total d'environ 12 à 13,5 millions de tokens selon les versions disponibles. Il constitue un corpus entièrement consacré à la science-fiction, mais n'est pas librement accessible, ce qui limite sa réutilisation dans des cadres ouverts ou reproductibles. PhraséoAnticipation-fr regroupe quant à lui des œuvres françaises de 1862 à 1939, issues du domaine public, pour un total d'environ 7 millions de tokens. Ce corpus, centré sur l'anticipation, n'est également pas librement accessible, rendant son utilisation impossible.

Enfin, le corpus Anticipation [9], issu du projet ANR de même nom, rassemble près de 2 100 textes parus entre 1860 et 1940, dont un échantillon représentatif de 474 récits a été retenu pour des analyses statistiques. Les œuvres sont décrites à travers un ensemble de métadonnées précises (cadres spatio-temporels, types de sociétés imaginaires, références scientifiques, informations éditoriales). Toutefois, ce corpus ne met pas à disposition les textes intégraux, ce qui empêche toute annotation linguistique fine et rend impossible l'étude des mots-fictions au sein du récit.

Ainsi, l'examen des ressources disponibles montre que si certains corpus offrent une couverture générique ou diachronique pertinente (PhraseoRomSF-fr, PhraseoAnticipation-fr, Anticipation), ils ne fournissent pas les conditions nécessaires à une étude des mots-fictions : absence de textes intégraux dans certains cas, absence d'annotations, restriction d'accès, volume insuffisant de textes francophones réellement rattachables à la science-fiction. Les corpus littéraires généraux (Chapitres, Project Gutenberg) ne permettent pas non plus de constituer de manière fiable une base de science-fiction spécialisée sans un

[6] Site officiel du projet : https://chapitres.hypotheses.org
[7] Site officiel du projet : https://www.gutenberg.org

important travail de tri manuel. En l'absence d'un corpus francophone libre, spécialisé en science-fiction et spécifiquement annoté pour les phénomènes linguistiques tels que les mots-fictions, la constitution d'un nouveau corpus apparaît nécessaire, et nous permettra d'amorcer non seulement des études linguistiques sur l'utilisation des mots-fictions, mais ouvrira aussi la voie à des applications de traitement automatique des langues ainsi que des exploitations relevant du domaine récent des humanités numériques computationnelles, ce que ne permettaient pas (directement) les corpus existants.

3 Construction du corpus

3.1 Méthodologie d'acquisition des données

Le genre de la science-fiction se manifeste à travers une diversité de supports tels que les romans, les livres de jeu de rôle, les bandes dessinées et romans graphiques, les nouvelles, les cartes illustrées ou encore les jeux vidéo. Afin d'éviter la mise en place de protocoles d'extraction et d'annotation hétérogènes dépendant de la nature multimodale des sources, nous avons volontairement restreint notre corpus aux supports strictement textuels : nouvelles, novellas et romans. Les sources illustrées (bandes dessinées, romans graphiques), cartographiques ou vidéoludiques auraient nécessité des chaînes de traitement séparés, incluant des techniques de vision par ordinateur ainsi qu'un protocole d'annotation spécifique pour les mots-fictions présents dans les images. Le périmètre textuel garantit ainsi l'homogénéité du traitement et la comparabilité des résultats.

Pour constituer ce corpus textuel, nous nous sommes appuyés sur le corpus Anticipation, qui regroupe 474 fiches pour des titres francophones publiés entre 1860 et 1940. Ce choix répond à deux contraintes majeures : disposer de récits littéraires en langue française et garantir le caractère libre de droits des œuvres. Parmi ces fiches, nous avons sélectionné uniquement les titres présentant des marqueurs de genres appartenant à la science-fiction, identifiés par les étiquettes « utopie », « dystopie », « post-apocalyptique », « inventions / innovations scientifiques », « société imaginaire » ou la présence d'« éléments scientifiques imaginaires ». Cette étape permet d'exclure les textes plus périphériques au genre, tels que les récits d'anticipation pure sans éléments spéculatifs. Par ailleurs, les quelques titres de bandes dessinées présents dans la base ont été retirés, car non conformes au périmètre textuel choisi. Ainsi, nous nous restreignons à une liste de titres déjà définie et approuvée par l'équipe de recherche du projet Anticipation. Par ailleurs, parmi les 246 récits finalement retenus, 10 sont presents dans les 25 romans catégorisés SF de Chapitres et 88 font partie des 109 livres de PhraseoAnticipation. Ces recoupements témoignent de la représentativité et de la légitimité des œuvres retenues pour le NovSFcorpus.

Pour la suite des étapes d'acquisition et d'annotation des données, nous adoptons la méthodologie établie pour le corpus LitBank [16], développé dans le cadre du projet Multilingual BookNLP de l'Université de Berkeley. Cette approche, déjà mise en œuvre pour d'autres ressources telles que LitBank Fr (voir Section 3.3), impose que le matériau textuel de départ soit constitué de fichiers au format brut (.txt) reproduisant le récit.

Pour cela, les œuvres retenues ont été récupérées sous forme de fichiers PDF (lorsqu'elles venaient de Gallica, Archive.org et du Projet Gutenberg) ou de fichiers TXT (lorsqu'elles venaient des autres sources) disponibles via différentes plateformes selon la répartition suivante : 213 textes proviennent de Gallica, 8 de Wikisource, 8 d'Archive.org, 5 d'Ebook gratuits, 4 d'ArcheoSF, 3 du Projet Gutenberg, 2 de la Bibliothèque de Gloubik,

2 de la Bibliothèque numérique romande, et 1 du Bouquineux[8]. Les liens exacts pour chaque récit ont été directement fournis par les fiches du projet Anticipation, ce qui a considérablement facilité la collecte des données. Ils sont par ailleurs référencés dans le fichier de métadonnées de NovSFcorpus[9].

Afin d'obtenir des versions textuelles exploitables, chaque PDF a été converti en images JPG, puis en fichiers texte via reconnaissance optique de caractères (*optical character recognition*, OCR). Nous avons utilisé Gemini 2.5 Flash [17], un modèle multimodal récent offrant des performances à l'état de l'art pour la lecture de documents numérisés en anglais [18]. Certains titres n'ont pas pu être intégrés : soit parce qu'ils étaient initialement publiés en romans-feuilletons, dont seules des portions fragmentaires sont disponibles, soit en raison d'une qualité de numérisation insuffisante ne permettant pas d'obtenir un OCR exploitable.

Une phase de nettoyage manuel a ensuite été réalisée sur l'ensemble des fichiers TXT afin de supprimer les éléments paratextuels indésirables issus des numérisations. Ont notamment été retirés : numéros de page, en-têtes et pieds-de-page, légendes d'illustrations, ainsi que les autres éléments récurrents dans les éditions anciennes souvent très illustrées. Les pages nettoyées ont par la suite été concaténées pour reconstituer chaque récit complet. À cette étape, les césures de fin de ligne ont été résolues afin de restaurer les mots artificiellement fragmentés, opération indispensable pour garantir la qualité des traitements linguistiques ultérieurs.

Afin d'évaluer automatiquement la qualité des OCR obtenus, chaque mot de chaque récit a été comparé au lexique fourni par spaCy[10], considéré ici comme ressource lexicale de référence pour la langue française en traitement automatique des langues. Les mots absents du lexique ont été comptabilisés comme inconnus, catégorie regroupant à la fois des erreurs d'OCR et des unités lexicales absentes des dictionnaires, notamment les mots-fictions. À ce stade, aucune distinction n'est possible entre ces deux causes d'erreurs. Un taux de mots reconnus a ainsi été calculé pour chaque texte, fournissant une approximation quantitative de la qualité théorique de l'OCR (sans préjuger de la qualité pratique, favorable aux mots-fictions correctement OCRisés mais inconnus du lexique). Les scores obtenus varient entre 96,88 % et 100 %, ce qui correspond à une OCRisation de très haute qualité et n'a conduit à l'exclusion d'aucun texte pour insuffisance de lisibilité.

Au terme de ce processus, nous obtenons le corpus de 246 récits francophones relevant de la science-fiction, publiés entre 1860 et 1940. Voici quelques exemples de titres inclus :

- Achille Eyraud, *Voyage à Vénus*, 1865
- Jules Verne, *Vingt mille lieues sous les mers*, 1869
- Didier de Chousy, *Ignis*, 1883
- Auguste de Villiers de l'Isle-Adam, *L'Eve future*, 1886
- Jules Lermina, *Le Secret des Zippelius*, 1892
- Camille Flammarion, *La Fin du monde*, 1894
- Jean Richepin, *Le dernier inventeur*, 1900
- J.-H. Rosny aîné, *La Mort de la Terre*, 1910
- Maurice Renard, *Les Mains d'Orlac*, 1920
- Léon Daudet, *Le Napus, Fléau de l'an 2227*, 1927

[8] Respectivement, les sites officiels : https://gallica.bnf.fr, https://fr.wikisource.org, https://archive.org, https://www.ebooksgratuits.com, http://archeosf.publie.net, https://www.gutenberg.org, https://livres.gloubik.info, https://ebooks-bnr.com, http://www.bouquineux.com

[9] https://github.com/mducos/FrenchSFcorpus/blob/main/src/metadata.csv

[10] Site officiel : https://github.com/explosion/spaCy

L'ensemble des titres, des noms d'auteurs, des dates de publication et du nombre de tokens est consigné dans un fichier de métadonnées au format CSV.

3.2 Méthodologie d'annotation des données

À partir des fichiers bruts des récits, nous appliquons une segmentation en phrases, suivie d'une tokenisation, d'une lemmatisation et d'un étiquetage morphosyntaxique réalisés avec SpaCy en suivant la méthodologie mise au point par [16]. Concernant les mots-fictions, ceux-ci ont été identifiés manuellement dans les récits puis consignés dans un fichier de métadonnées associé à chaque récit[11]. Voici la méthodologie d'annotation précise :

- Les syntagmes décrivant les mots-fictions sont annotés, contrairement aux périphrases en décrivant un. Ainsi, *almanach électrique* constitue un mot-fiction car il forme une unité lexicale autonome, tandis qu'une description telle que « Enfin, pour préserver par tous les moyens la coque elle-même de la pression éventuelle des glaces, on l'enveloppa d'un berceau d'acier dont les bandes, reliées entre elles par une série de croix de Saint-André et étirées dans des filières en croix, reposaient sur des poutrelles également de fer, encastrées dans une gangue de bois. Ainsi soutenu, le navire ne devait rien redouter de la pression sur sa quille ou ses flancs. » (*Une Française au Pôle nord*, Maël, 1893) pour décrire un berceau protégeant un bateau de la pression des glaces n'en est pas un, même si elle décrit un objet fictif.
- Une expression est considérée non-lexicalisée lorsqu'elle est absente du Wiktionnaire[12], dictionnaire retenu pour sa couverture large incluant les néologismes et termes non encore intégrés aux dictionnaires officiels, ainsi que pour son utilisation dans des projets traitement des expressions polylexicales [20, 21]. Font exception les termes présents dans le Wiktionnaire dont l'étymologie atteste d'une origine fictionnelle correspondant à la date de publication du récit ou lorsque la définition est étiquetée « Science-fiction », comme pour l'entrée *ferromagnétal*[13].
- Sont également retenus comme mots-fictions les termes dont la forme est lexicalisée mais dont la sémantique est totalement nouvelle par rapport aux définitions attestées (par exemple *tube* désignant un tunnel souterrain reliant les continents entre eux dans *Le vingtième siècle*, Robida, 1883).
- Les métaphores, expressions (comme « se mettre sur son trente-et-un »), jeux de mots et autres figures de style ne relèvent pas de mots-fictions, de même que les mots appartenant à une autre langue que le français (langue inventée comprise).
- Les noms propres de personnages, de lieux ou d'organisations inventées, car ils relèvent des entités nommées et non de concepts ou d'objets servant à concevoir le monde fictionnel [13]. Ainsi, *ville aérienne* est un mot-fiction, mais le nom donné à cette ville ne l'est pas.
- Les adjectifs et substantifs dérivés de noms propres constituent des mots-fictions lorsqu'ils sont formés par dérivation et désignent un concept ou une caractéristique du monde fictionnel [13], comme *andréïdien*.
- Lorsqu'un adjectif seul représente un mot-fiction, deux unités sont annotées séparément : l'adjectif seul d'une part, et le syntagme nom-adjectif d'autre part, ce

[11] Ce fichier est accessible à l'adresse : https://github.com/mducos/FrenchSFcorpus/blob/main/src/title2novum.json
[12] Site officiel : https://fr.wiktionary.org
[13] Section étymologie de *ferromagnétal* dans le Wiktionnaire : « 1910. Création de de J.-H. Rosny aîné. De *ferromagnétique*. » (https://fr.wiktionary.org/wiki/ferromagn%C3%A9tal)

dernier formant un mot-fiction à part entière. Cette double annotation se justifie par le fait qu'un même adjectif peut, selon le nom auquel il se rapporte, désigner des référents fictifs distincts : ainsi, l'adjectif *napusien* associé à *épidémie* forme le mot-fiction *épidémie napusienne*, désignant un phénomène pathologique fictif, tandis que *savate napusienne* désigne un tout autre objet fictif. L'adjectif seul constitue donc un mot-fiction en tant qu'unité dérivationnelle porteuse d'un sens fictionnel, et chaque syntagme nom-adjectif constitue un mot-fiction supplémentaire et distinct dont le référent est déterminé par la combinaison des deux unités.

- Les mots-fictions sont systématiquement fléchis au singulier dans les métadonnées afin de permettre l'annotation automatique de toutes leurs occurrences fléchies dans les textes.

Toujours en suivant la méthodologie de constitution du corpus LitBank, et plus précisément les procédures d'annotation des entités [16], les mots-fictions collectés manuellement dans le fichier de métadonnées ont ensuite été annotés automatiquement dans l'ensemble des récits, pour chacune de leurs occurrences. Les données d'entités, dans notre cas uniquement les mots-fictions, les autres catégories d'entités (personnages, organisations, lieux, etc.) étant traitées ultérieurement, sont mises au format brat standoff et au format tabulaire décrit par [22]. La correspondance entre le mot-fiction et ses occurrences dans le texte s'appuie à la fois sur le lemme produit par spaCy et sur la forme brute du token, afin de pallier les erreurs de lemmatisation du modèle sur les mot-fiction. Chaque ligne comporte plusieurs colonnes séparées par une tabulation de la forme :

word lemma label1 label2 … labelN

Le nombre de labels (N) correspond au niveau maximum d'imbrication présent dans le jeu de données, augmenté de 1. Dans notre corpus, ce niveau maximal est de 2. De plus, chaque phrase est séparée par une ligne vide dans ce fichier d'annotation. Par exemple, pour les deux phrases suivantes, qui contiennent trois mots-fictions, *homme lunaire*, *femme lunaire* et *papillon à face humaine*, le format attendu est :

Ils	il	O	O	O
virent	voir	O	O	O
un	un	O	O	O
homme	homme	B-NOV	O	O
et	et	O	O	O
une	un	O	O	O
femme	femme	O	B-NOV	O
lunaires	lunaire	I-NOV	I-NOV	O
.	.	O	O	O

Avec B pour *beginning*, I pour *inside* et O pour *outside*.

Finalement, chaque récit est associé à quatre fichiers, dont voici un exemple avec *Autour de la Lune* de Jules Verne :

- *Verne_AutourDeLaLune_1870.txt* : texte original ;
- *Verne_AutourDeLaLune_1870_sent.txt* : segmentation en phrases, avec une phrase par ligne ;
- *Verne_AutourDeLaLune_1870.tsv* : annotation au format BIO des mots-fictions après tokenisation ;

- *Verne_AutourDeLaLune_1870.ann* : annotation des mots-fictions en span avec leurs positions en caractères.

S'y ajoute le fichier de métadonnées au format CSV[14], avec les auteurs, les titres correspondants et leur date de parution, le nombre de token pour chaque récit, et les liens de téléchargement vers les plateformes décrites dans la Section 3.1.

Voici un ensemble d'exemples de mots-fictions relevés dans le corpus :

- train souterrain (*Voyage à Vénus*, Eyraud, 1865)
- columbiad (*Autour de la Lune*, Verne, 1870)
- phono-annonceur à clavier (*Le Vingtième siècle*, Robida, 1883)
- appareil à mûrir les bébés (*La Vie électrique*, Rameau, 1887)
- fleur lunaire (*Voyage dans la Lune avant 1900*, Avray, 1892)
- voiture électrique (*La mort de Paris*, Gallet, 1892)
- anthroposaure (*Le Peuple du pôle*, Derennes, 1907)
- téléphone sans fil (*Les Gratteurs de ciel*, Boussenard, 1907)
- oeil pataphysique (*Gestes et opinions du docteur Faustroll*, Jarry, 1911)
- almanach électrique (*Les Plaisirs de la campagne*, Mac Orlan, 1912)
- appareil héliophorique (*Utopie des îles bienheureuses dans le Pacifique en l'an 1980*, Masson, 1921)
- cinébouquin (*Le Napus : fléau de l'an 2227*, Daudet, 1927)

Ces exemples montrent qu'un mot-fiction peut être une expression polylexicale ou un lexème simple.

Cependant, bien que ces formes présentent toutes un signifiant nouveau, il ne suffit pas d'identifier les termes non lexicalisés (ou hors-vocabulaire d'un point de vue computationnel) pour détecter automatiquement l'ensemble des mots-fictions. Ainsi, *cinébouquin* sera correctement repéré comme terme hors-vocabulaire, mais *almanach électrique* ou *appareil héliophorique* ne le seront pas : dans ce dernier cas, seul *héliophorique* apparaît hors vocabulaire, alors que le mot-fiction repose sur l'expression complète.

Ces trois exemples illustrent les différentes configurations permettant d'établir une typologie des mots-fictions : le mot-fiction complet est hors vocabulaire ; seule une partie du mot-fiction est hors vocabulaire ; l'ensemble du mot-fiction est lexicalisé. L'ensemble des types possibles de mot-fiction peut ainsi être défini en combinant ces trois situations : non-lexicalisation, lexicalisation partielle ou lexicalisation complète des composants. Le Tableau 1 récapitule l'ensemble des cas de figure.

Dans le fichier CSV contenant les annotations manuelles, chaque mot-fiction a été associé au numéro de son type. Cette information supplémentaire permet de conduire des analyses plus fines sur la répartition des différentes catégories de mots-fictions au sein du corpus.

Table 1. Typologie des mots-fictions.

Mot(s) existant(s)	Expression polylexicale existante	Type d'expression	Exemples
Oui	Oui	Expression du lexique commun	éclairage artificiel infomatique train à vapeur

[14] https://github.com/mducos/FrenchSFcorpus/blob/main/src/metadata.csv

	Non	Mot-fiction (classe 1)	almanach électrique appareil à mûrir les bébés fleur lunaire
Non	Oui	*Cas impossible*	-
	Non	Mot-fiction (classe 2)	cinébouquin anthroposaure columbiad
Partiellement	Oui	*Cas impossible*	-
	Non	Mot-fiction (classe 3)	phono-annonceur à clavier oeil pataphysique appareil héliophorique
Oui ou non ou partiellement	Expression inexistante au moment de l'écriture, mais existante aujourd'hui	Mot-fiction lexicalisé (classe 4)	voiture électrique téléphone sans fil train souterrain

4 Description et analyse du corpus

4.1 Analyse des types de mots-fictions

Dans nos annotations manuelles, nous avons relevé 945 types de mots-fictions au total[15], dont 708 types polylexicaux et 237 types simples, avec une moyenne de 2,34 tokens par mot-fiction. Le corpus sur lequel repose cette annotation comprend 246 récits pour un total de 10 255 271 tokens et 46 338 573 signes (espaces non comprises) : 114 nouvelles (jusqu'à 17 000 mots), 29 novellas (entre 17 000 et 40 000 mots) et 103 romans (au-delà de 40 000 mots). La Figure 1 présente la répartition de ces récits selon leur date de publication.

Les annotations du corpus portant sur les mots-fictions, les statistiques y font principalement référence. Ainsi, sur les 246 récits, 75 d'entre eux ne contiennent aucun mot-fiction annoté (au sens de syntagme mais possèdent malgré ça des périphrases non-annotées décrivant un mot-fiction), tandis que 171 récits en comportent au moins un. Le corpus totalise 945 types de mots-fictions, tous types confondus, soit une moyenne de 3,84 mots-fictions par récit. Si l'on exclut les récits sans mot-fiction, cette moyenne augmente à 5,53 mots-fictions par récit, ce qui souligne une concentration marquée des créations lexicales dans une partie seulement des textes.

La distribution par type montre une prédominance nette des mots-fictions de classe 1 (570 types de mots-fictions), suivis des mots-fictions de classe 2 (258 types). Les classes 3 (55 types) et 4 (62 types) apparaissent beaucoup plus marginalement sans pour autant être non-significatifs car 31 récits contiennent des mots-fictions de type 3 et 37 en contiennent de type 4, comme le montre la Figure 2.

[15] Le fichier comportant les types de mots-fictions par récit, accompagné de leur classe et de leur distribution morpho-syntaxique est disponible à cette adresse : https://github.com/mducos/FrenchSFcorpus/blob/main/src/title2novum.json

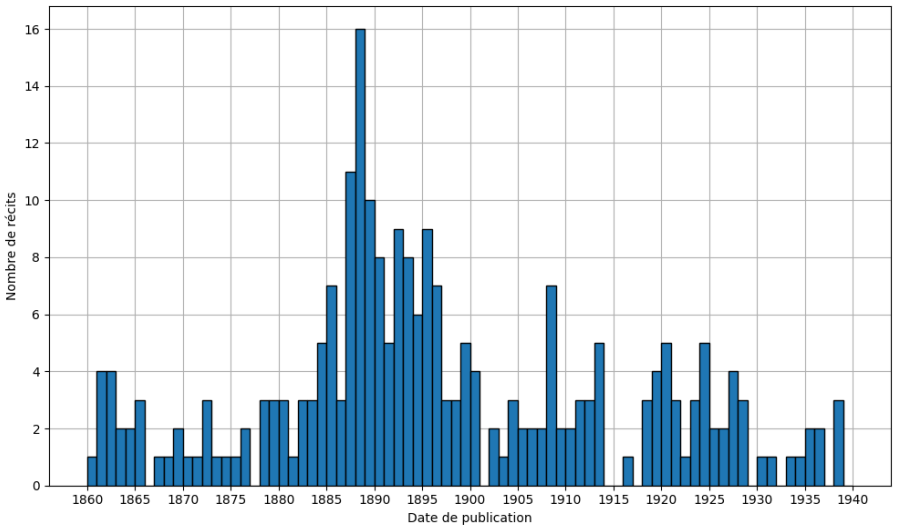

Fig. 1. Distribution des dates de publication des récits.

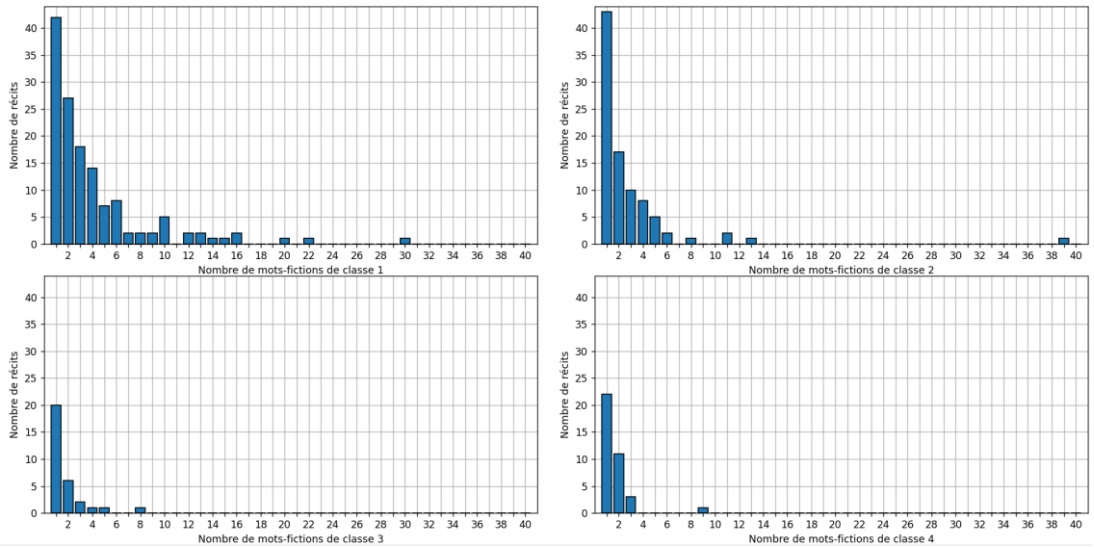

Fig. 2. Distribution du nombre de mots-fictions par récit selon leur type.

Afin d'examiner la distribution morpho-syntaxique des mots-fictions, la librairie Stanza[16] a permis d'assigner à chaque mot-fiction une séquence de catégories morpho-syntaxiques correspondant à chaque terme de l'expression. À partir de ces annotations, nous avons produit une visualisation de la distribution des patrons morpho-syntaxiques observés dans l'ensemble des types de mots-fictions du corpus.

Les résultats montrent une prédominance nette des structures NOUN-ADJ, avec 319 types de mots-fictions (par exemple *almanach électrique* ou *affichage stellaire*). Viennent ensuite les NOUN simples, avec 202 types (par exemple *anthroposaure* ou *atomologie*), qui correspondent majoritairement à des créations lexicales à un seul lexème. Les structures

[16] Site officiel : https://stanfordnlp.github.io/stanza/

plus complexes de type NOUN-ADP-NOUN comptent 64 types (par exemple *gratteur de ciel* ou *œufs en celluloïd*) et celles avec une ponctuation comme NOUN-PUNCT-NOUN (par exemple *photo-téléphonographe* ou *auto-compositrice*) ou NOUN-ADP-PUNCT-NOUN (par exemple *fleuve d'astéroïde* ou *telephone sans-fil*) sont respectivement au nombre de 56 et 25. La Figure 3 montre la distribution des dix combinaisons de catégories morpho-syntaxiques les plus fréquentes.

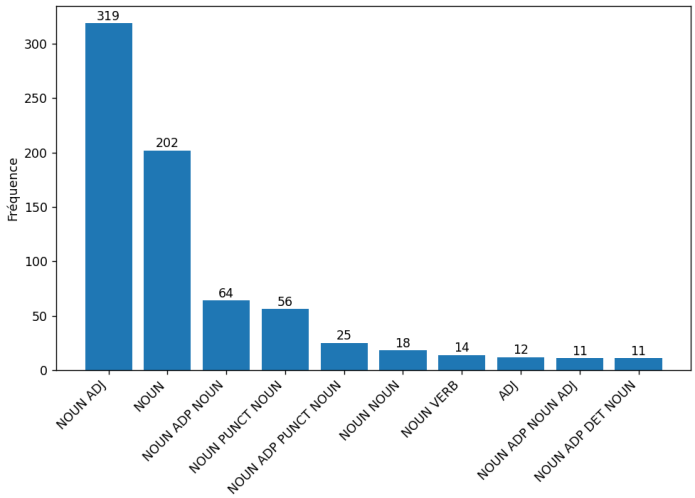

Fig. 3. Distribution des catégories morpho-syntaxiques de chaque terme composant le mot-fiction, pour chacun des 945 types de mot-fiction différents. Seules les dix combinaisons les plus fréquentes sont représentées dans cette figure

Cette diversité dans la distribution morpho-syntaxique des mots-fictions montre que leur construction s'appuie sur des structures variées. Cette distribution était attendue par la façon dont les annotations ont été conduites.

4.2 Evaluation de l'annotation des occurrences des mots-fictions

L'annotation automatique des occurrences de mots-fictions à partir de la liste établie manuellement a permis d'identifier 3 931 occurrences de mots-fictions réparties sur 3 537 phrases, sur un total de 597 697 phrases dans l'ensemble du corpus (0,6% des phrases possédant un mot-fiction), ce qui reflète le caractère très rare des mots-fictions dans les textes.

Afin d'évaluer la qualité des annotations produites, un échantillonnage a été réalisé à partir de l'ensemble des phrases du corpus. Deux sous-échantillons de 500 phrases chacun ont ainsi été constitués. Le premier sous-échantillon est composé de 500 phrases contenant au moins une annotation NOV (soit 559 mots-fictions et 932 tokens associés, étiquettes B-NOV et I-NOV confondues). L'objectif est de vérifier la justesse des annotations produites automatiquement, cette première évaluation correspondant à la précision. Le second sous-échantillon est composé d'un autre ensemble de 500 phrases ne contenant aucune annotation NOV. L'objectif est cette fois-ci d'estimer le rappel, c'est-à-dire de vérifier si des mots-fictions ont été omis par le système d'annotation.

L'évaluation manuelle du premier sous-échantillon révèle 11 erreurs sur 500 phrases, soit une précision de 98,0% (11 mots-fictions mal annotés sur 559). Ces erreurs sont imputables à la phase d'annotation automatique et non à l'annotation manuelle des types. Elles sont dues à des ambiguïtés catégorielles que le système automatique ne résout pas :

par exemple, le syntagme *être artificiel* constitue un mot-fiction lorsque *être* est employé comme nom commun, mais le pipeline d'annotation ne distingue pas cette occurrence des cas où *être* est un verbe, produisant ainsi de faux positifs.

Concernant le second échantillon, son évaluation manuelle révèle 3 erreurs sur 500 phrases. Contrairement aux erreurs du premier sous-échantillon, celles-ci ne sont pas dues à la phase automatique mais à des omissions lors de l'annotation manuelle initiale des types, ce qui signifie que ces mots-fictions n'ont jamais été intégrés à la liste de référence et n'ont donc pas pu être détectés automatiquement.

Ces résultats indiquent que le pipeline d'annotation présente une précision et un rappel très satisfaisants (98,0% de précision et 99,4% de rappel). Les erreurs de précision, peu nombreuses, sont concentrées sur des cas d'ambiguïté catégorielle que pourrait résoudre une prise en compte du POS assigné par spaCy lors de la phase automatique. Les erreurs de rappel, quant à elles, relèvent de l'annotation manuelle et pourraient être réduites par l'ajout des mots-fictions ainsi découverts à la liste des types de mots-fictions par récit.

Le corpus actuel tel que téléchargeable sur le github présente les corrections manuelles sur les types non-annotés, autrement dit les 3 erreurs impactant le rappel.

5 Conclusion et perspectives

Cet article a montré la méthodologie de la constitution d'un corpus de récits littéraires. La construction systématique de ce corpus a permis d'identifier et de catégoriser les mots-fictions, offrant une base structurée pour toute étude quantitative ou qualitative ultérieure. Le corpus ainsi constitué constitue un outil précieux pour étudier les mots-fictions, ces éléments qui introduisent des innovations dans les récits. Leur suivi permet non seulement de mieux comprendre la créativité narrative des auteurs, mais aussi d'observer comment certaines idées fictionnelles peuvent précéder des technologies réelles, à l'image du *téléphone sans fil* [23].

Pour le corpus étudié, une perspective majeure est le développement d'un outil de détection automatique de mots-fictions. Une telle approche permettrait de repérer plus rapidement les inventions linguistiques imaginaires et de suivre leur évolution dans la littérature, ouvrant ainsi la voie à des études comparatives et à des applications interdisciplinaires entre littérature et linguistique.

References

1. S. Bréan, La Science-fiction en France : Théorie et histoire d'une littérature (Presses Université Paris-Sorbonne, Paris, 2012)
2. S. Bérard et M. Uhl, La science-fiction : de la perspective à la prospective. Les Cahiers de la CRSDD. **8** (2012)
3. U. Eco, Lector in fabula. Le rôle du lecteur ou la coopération interprétative dans les textes narratifs (Librairie générale française, Paris, 1985)
4. D. Suvin, Metamorphoses of Science Fiction: On the Poetics and History of a Literary Genre (Yale University Press, New Haven, 1979)
5. M. Angenot, La Parole Pamphlétaire. Contribution à la typologie des discours modernes (Payot, Paris, 1982)
6. N. Vas Deyres, Ces Français qui ont écrit demain : utopie, anticipation et science-fiction au XX^{e} siècle (Honoré Champion, Paris, 2013)
7. A. Evans, La science-fiction fantastique de Maurice Renard. Res Futurae. **11** (2018) https://doi.org/10.4000/resf.1439

8. I. Langlet, La science-fiction, lecture et poétique d'un genre littéraire (Armand Colin, Paris, 2006)
9. C. Barel-Moisan, Le Roman des possibles. L'anticipation dans l'espace médiatique francophone (1860-1940) (Presses Universitaires de Montréal, Montréal, 2019)
10. J. Pruvost et J.-F. Sablayrolles, Les néologismes (Presses universitaires de France, Paris, 2019)
11. J. Semprun, Défense et illustration de la novlangue française (Editions de l'Encyclopédie des Nuisances, Paris, 2005)
12. R. Saint-Gelais, L'Empire du pseudo. Modernités de la science-fiction (Nota Bene, Québec, 2005)
13. A. Ray, Traduire les termes du futur : Analyse du traitement des termes-fictions dans la traduction de l'anglais au français de la littérature de science-fiction. Thèse de doctorat, Université d'Orléans (2019)
14. P. Gindre, La formation des néologismes dans la littérature de science-fiction d'expression anglais contemporaine. Thèse de doctorat, Université de Franche-Comté (1999)
15. S. Diwersy, La phraséologie du roman contemporain dans les corpus et les applications de la PhraseoBase. Corpus. **22** (2021) https://doi.org/10.4000/corpus.6101
16. D. Bamman, An Annotated Dataset of Coreference in English Literature, in Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France, mai (2020)
17. Gemini team, Gemini: A Family of Highly Capable Multimodal Models. ArXiv preprint (2025). https://doi.org/10.48550/arXiv.2312.11805
18. F. Ling, OCRBench v2: An Improved Benchmark for Evaluating Large Multimodal Models on Visual Text Localization and Reasoning. ArXiv preprint (2025) https://doi.org/10.48550/arXiv.2501.00321
19. D. Bamman, An Annotated Dataset of Literary Entities, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Minnesota, États-Unis, juin (2019). https://doi.org/10.18653/v1/n19-1220
20. Krstev, C., Progress in SR-ELEXIS Semantic Annotation: Focusing on Multiword Expressions, Named Entities, and Sense Repository. In 3rd General Meeting Hungarian Research Centre for Linguistics, Budapest, Hungary 29-30 January 2025. European Cooperation in Science and Technology (2025)
21. Überrück-Fries, T., Savary, A., & Dryjańska, A., Sailing through multiword expression identification with Wiktionary and Linguse: A case study of language learning. In Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning (2024)
22. M. Ju, A Neural Layered Model for Nested Named Entity Recognition, Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Nouvelle Orléans, Louisiane, États-Unis, juin (2018). https://doi.org/10.18653/v1/n18-1131
23. E. Picholle, Les défis du novum, de la science-fiction à l'histoire des idées scientifiques. Espace et temps. **4** (2018)