Analyse qualitative et quantitative des “ hallucinations ” générées automatiquement dans un corpus de reformulations médicales

Ioana Buhnila, Georgeta Cislaru, Amalia Todirascu

Analyse qualitative et quantitative des « hallucinations » générées automatiquement dans un corpus de reformulations médicales

Ioana Buhnila[1,*], *Georgeta* Cislaru[2] et *Amalia* Todirascu[3]

[1]ATILF UMR 7118 (CNRS - Université de Lorraine)
[2]MoDyCo UMR 7114 (Université Paris Nanterre)
[3]LiLPa UR 1339 (Université de Strasbourg)

[*]ioana.buhnila@univ-lorraine.fr

Résumé. Notre étude porte sur les « hallucinations », des productions langagières générées par des outils d'intelligence artificielle de type générateurs de textes, productions qui ne correspondent pas à ce qu'il est attendu de l'outil. Nous testons l'hypothèse selon laquelle il est possible de discerner des patrons langagiers dans ces générations inadéquates. Nous menons des analyses quantitatives et qualitatives des données, selon plusieurs entrées : le degré d'adéquation grammaticale et sémantique des séquences générées, les relations sémantiques, les fonctions sémantico-pragmatiques et les discrépances combinatoires. Nos analyses montrent que les outils de génération textuelle procèdent à de généralisations abusives en mettant en exergue des patrons dont la portée n'est pas validée par l'usage. D'un point de vue informatique, les « hallucinations » soulèvent des questions quant au paramétrage des modèles langagiers exploités par les réseaux neuronaux et la génération statistique. D'un point de vue linguistique, nos observations soulèvent la question de l'interface entre les usages purement linguistiques et leurs différents contextes sur le terrain des pratiques langagières qui ancrent ces patterns dans l'usage.

1 Introduction

Notre étude porte sur ce que l'on appelle souvent les *hallucinations* de l'IA. Il s'agit de productions langagières générées par des outils d'intelligence artificielle de type générateurs de textes : les séquences qui ne correspondent pas à ce qu'il est attendu en matière de production sont appelées des hallucinations.

Ces séquences formellement ou sémantiquement non conformes soulèvent un certain nombre de questions quant à l'exploitation des modèles langagiers par les outils de génération automatique et, plus particulièrement, l'apprentissage des structures utilisées fréquemment dans les textes (patrons morphosyntaxiques ou sémantico-relationnels). Nous allons tester l'hypothèse selon laquelle il est possible de discerner des régularités et donc des patrons langagiers dans les générations inadéquates.

Pour ce faire, nous analysons les aberrations langagières issues de la génération automatique de reformulations médicales en français réalisées par (Buhnila, 2023). La section 2 discute de la notion d'hallucination et propose une brève revue de la littérature des travaux traitant de ce phénomène. Nous présentons les données dans la section 3, tandis que la section 4 expose l'approche méthodologique adoptée. La section 5 détaille les analyses quantitatives et qualitatives des données, selon plusieurs entrées : degré d'adéquation grammaticale et sémantique des séquences générées, relations lexicales, fonctions sémantico-pragmatiques, discrépances combinatoires.

2 Contexte de l'étude

Le terme-même de « hallucination » est remis en cause par une série de publications récentes issues notamment du domaine médical. Ainsi, s'insurgeant contre ce qu'ils considèrent une extension sémantique abusive d'un terme médical, Østergaard et Nielbo (2023) insistent sur le caractère de « fausse réponse » des générations langagières non conformes, tandis que Hatem et al. (2023) utilisent le terme d'« affabulation » (*confabulation* en anglais). Dans son éditorial, Emsley (2023) parle de fabrications et

falsifications et soulève notamment la question des références bibliographiques inventées – qui, donc, n'existent pas – susceptibles de remettre en question les études qui y feraient appel (voir aussi Alkaissi et McFarlane, 2023 ; Athaluri et al., 2023). Ces mises en garde, qui soulèvent de nombreuses questions éthiques afférentes à l'usage des outils IA, mettent le spécialiste en sciences du langage face à une problématique spécifique concernant ce que l'on peut appeler du langage artificiel, les modèles langagiers sous-jacents et la nature des séquences ne répondant pas à des conditions de conformité formelle ou de vérité sémantique.

La nécessité de trouver des solutions pour corriger les productions des outils IA a donné lieu à de nombreux travaux s'intéressant à l'analyse et à la classification des « hallucinations » ; une revue détaillée est proposée dans Ye et al. (2023). Des outils d'identification sont également mis en place, comme Ne-Mo Guardrails (Nvidia). Généralement, la littérature s'intéressant à la typologie n'aborde pas les productions du point de vue des structures langagières, mais du point de vue des tâches demandées à l'outil, comme la traduction automatique, les questions-réponses, les résumés, etc. (Bruno et al., 2023 ; Zhang et al., 2023).

Or, les modèles langagiers et algorithmes qui sous-tendent les générateurs de textes sont tributaires d'une série de schématisations préalables susceptibles de déterminer le résultat produit. Se situant dans une perspective vygotskienne, De Castro et Zona (2022) soulignent l'intervention des stéréotypes culturels dans l'apprentissage des algorithmes et le biais de généralisation qu'ils induisent.

Le rôle des patterns langagiers est discuté par Durt et al. (2023), selon lesquels les outils AI modélisent et reproduisent les schémas d'usages linguistiques et non des schémas mentaux, les relations distributionnelles ne rendant compte que d'un aspect des relations sémantiques. Les auteurs portent un regard critique sur les corpus d'apprentissage issus de données écrites, qui ne représentent que partiellement l'usage, et de ressources médiatiques et numériques, dont ils mentionnent le pouvoir amplificateur des répétitions, préjugés et clichés :

> Unoriginal text can furthermore appear human-like for an embarrassing reason: The mindless repetition and reassociation of patterns is by no means limited to machines. Human thinking, speaking, and writing are often much less authentic than we would like to admit. As Heidegger famously observed, much of what people do is done because that is how "one" does things (Heidegger, 2010). (Durt et al., 2023 : 11)[i]

Nous gardons donc à l'esprit que le sens n'est valable qu'en usage et que les séquences que nous observons ne sont ni des hallucinations à proprement parler, ni des unités porteuses de sens, mais des émanations d'une certaine modélisation des patterns langagiers que nous nous proposons de scruter à travers nos analyses.

3 Les données : génération automatique de reformulations dans le domaine médical

Nos données sont issues d'un projet plus large s'intéressant à l'analyse linguistique des reformulations présentes dans des corpus scientifiques et grand public du domaine médical (Buhnila, 2023). Des expériences d'apprentissage automatique par réseaux de neurones ont été réalisées en utilisant l'architecture neuronale contradictoire APT (*Adversarial Paraphrasing Task*) (Nighojkar et Licato, 2021) et le modèle de langue générale disponible pour le français T5 (*Text to Text Transformer*) (Raffel et al., 2020)[ii], adapté pour le domaine médical.

L'architecture neuronale APT (Nighojkar et Licato, 2021) a été initialement conçue pour générer des paraphrases phrastiques et sous-phrastiques en anglais non spécialisé. Le modèle vise à identifier le sens général des paraphrases phrastiques ou sous-phrastiques, au-delà d'un seul mot à la fois. La particularité d'APT est qu'elle permet de générer des paraphrases avec des significations équivalentes mais présentant des variations lexicales et syntaxiques. En effet, contrairement à d'autres systèmes de génération de paraphrase disponibles en anglais (Witeveen et Andrews, 2019 ; Palivela, 2021 ; Vernikos et Popescu-Belis, 2024), le système APT privilégie la génération de paraphrases dont la forme est très différente de l'original, mais le sens reste assez proche. Pour trouver ce type de paraphrases, deux scores sont exploités :

i) le score MI (implication mutuelle) permet de vérifier si les paraphrases sont mutuellement implicites et sémantiquement équivalentes (c'est-à-dire qu'il est possible de déduire la paraphrase partant de la phrase originale et vice-versa), à l'aide d'un modèle de langue RoBERTa (Nie et al, 2020), adapté pour la tâche d'inférence ; ii) le score BLEURT (Sellam et al., 2020), basé sur les plongements des mots du modèle de langue BERT (Devlin et al., 2018), évalue les différences lexicales et syntaxiques des paraphrases générées automatiquement. Un score BLEURT supérieur à 0,5 indique que les paraphrases sont trop similaires. Ainsi, les paraphrases qui sont retenues sont celle qui ont un score MI similaire, mais un score BLEURT réduit (inférieur à 0,5).

Les paramètres d'APT ont été modifiés afin de rendre le modèle opérationnel en français dans le champ du discours médical (Buhnila, 2023) et adapté pour une tâche de génération de paraphrase sous-phrastiques. Peu de modèles de langue sont disponibles en français pour la génération de textes (GPT-4 (OpenAI) et Claude (Anthropic), des modèles dont le code source est privé, ou BARThez (Eddine et al, 2020), en accès libre). En général, ces modèles proposent des paraphrases phrastiques voir des textes complets, ce qui n'est pas notre objectif. Par ailleurs, dans notre cas, une tâche différente a été ciblée : la génération des paraphrases phrastiques a été remplacée avec la génération des *reformulations sous-phrastiques,* telles que définies dans la sous-section suivante.

3.1 La reformulation sous-phrastique médicale

Selon Buhnila (2023), la *reformulation sous-phrastique médicale* est un processus de transformation du discours avec l'objectif d'expliquer, de simplifier ou de clarifier une notion médicale ou un syntagme tout en restant dans l'empan de la même phrase. Ces reformulations peuvent prendre la forme d'une explication, d'une définition, d'une paraphrase (syntagmes synonymiques) ou même d'une exemplification. Un corpus de reformulations sous-phrastiques médicales constitué semi-automatiquement a été utilisé pour la génération automatique des prédictions médicales. Nous le présentons ci-dessous.

3.2 RefoMed : corpus de reformulations médicales

RefoMed (*Reformulations Médicales*)[iii] (Buhnila, 2023) est un corpus de reformulations sous-phrastiques médicales constitué de 8626 paires de termes médicaux – reformulations médicales en français[iv]. Les paires ont été utilisées pour l'adaptation d'APT pour la tâche de génération de reformulations. Les termes médicaux et leurs reformulations ont été extraits de corpus comparables de textes scientifiques et de textes de vulgarisation du domaine de la médecine : les corpus ClassYN (Todirascu et al., 2012) et CLEAR Cochrane (Grabar et Cardon, 2018).

La méthode semi-automatique de construction du corpus comporte plusieurs étapes de traitement automatique : sélection automatique des phrases contenant des termes médicaux (à l'aide de SIFR-BioPortal (Tchechmedjiev et al., 2018) et des marqueurs de reformulation identifiés dans la littérature (Fuchs, 1982 ; Grabar et Eshkol-Taravella, 2017 ; Vassiliadou, 2020), suivie d'annotations manuelles et d'une validation manuelle, et un calcul du score inter-annotateur Kappa. L'annotation manuelle valide si le marqueur introduit une reformulation pour le terme identifié automatiquement. De même, suivant Grabar et Eshkol-Taravella (2017), ont été annotées les relations lexicales s'établissant entre le terme et sa reformulation (hyperonymie, hyponymie, synonymie, méronymie) et les fonctions sémantico-pragmatiques associées (définition, exemplification, paraphrase, explication) à la reformulation. Le besoin d'automatiser la tâche d'identification et de génération des reformulations sous-phrastiques nous a orienté vers la prise en compte des marques linguistiques repérables par un algorithme, comme les marqueurs de reformulation (Buhnila, 2022).

Dans le discours, les reformulations sont introduites par des marqueurs spécifiques, de type « autrement dit », « signifie », « c'est-à-dire », etc. (Fuchs, 1982 ; Gülich et Kotchi, 1987 ; Grabar et Eshkol-Taravella, 2017 ; Vassiliadou, 2020). Ces éléments lexicaux, grammaticaux (*tels que, par exemple*), ou orthographiques (parenthèses, double points, virgule) sont d'une grande importance dans l'identification de la reformulation sous-phrastique (Buhnila, 2022). Néanmoins, certaines reformulations ne comportent

pas de marquage (Vassiliadou, 2020) : bien que faciles à identifier par un lecteur humain, elles présentent des difficultés pour la machine.

3.3 Génération automatique des reformulations

La génération automatique a été réalisée à l'aide de l'architecture neuronale contradictoire APT (Nighojkar et Licato, 2021) et du modèle de langue T5 (Raffel et al., 2020). Nous avons entrainé le modèle de langue générale sur notre tâche spécifique avec les 8146 paires de termes (= reformulations correctes) du corpus RefoMed, dont 480 paires ont constitué le jeu de données de test et le reste le jeu de données d'entraînement du modèle de langue. Nous avons utilisé le jeu d'entraînement pour *finetuner* notre système. Le *finetuning* est une méthode d'apprentissage supervisé qui permet d'adapter le modèle T5 à la tâche de génération de reformulations sous-phrastiques médicales.

Suivant (Nighojkar et Licato, 2021), nous avons utilisé une fonction de perte (*loss*) spécifique à la tâche : cette fonction calcule la *perte d'entropie croisée* obtenue sur chaque lot (en anglais *batch*), divisée par une valeur *gain* qui permet d'évaluer la diversité lexicale et syntaxique de la paraphrase (si le modèle a réussi à obtenir une valeur de 1 pour ce batch, alors la valeur est 1, sinon, on garde la valeur calculée de la perte d'entropie croisée). La formule appliquée pour calculer le gain est :

$$\text{gain} = \text{mi}/(1+e^{5*\text{bleurt}})^2$$

ou mi = 1 si le terme et la paraphrase sont sémantiquement équivalents ; la valeur maximale du gain est maximum 1 et le gain est 0 si les phrases ne sont pas sémantiquement équivalents ou si BLEURT est > 0,5. Le calcul de MI et de BLEURT s'appuie sur CamemBERT (Martin et al, 2020).

Tableau 1. Valeurs de la fonction de loss obtenues pendant le finetuning du modèle de langue T5 avec les paires terme médical – reformulation correcte du corpus RefoMed.

epoch	Loss
0	2,14
1	1,27
2	0,96
3	0,89
4	0,74
5	0,62

L'apprentissage a été réalisé sur 15 lots (en anglais *batch*), nous avons appliqué 6 itérations (en anglais epochs) et le pas pour modifier les poids est 0.01 et le taux d'apprentissage (3e-4). Après 4 itérations, le plateau est atteint. Nous utilisons *Adam optimizer* pour le calcul (adam_epsilon=1e-8). Nous avons testé notre méthode sur 480 termes choisis aléatoirement, en générant entre 1 et 5 prédictions pour chaque terme de la liste de test. Nous présentons une évaluation avec des métriques de similarité sémantique dans l'**Annexe 1**. Dans la section suivante, nous proposons une analyse qualitative et quantitative des 2266 prédictions générées automatiquement.

4 Méthode

Pour étudier les générations erronées et en établir une typologie, nous avons d'abord réalisé une extraction de n-grammes disponibles sur le corpus de résultats obtenus par la génération automatique de reformulations, sur la plateforme SketchEngine (Kilgarriff et al., 2014) (section 4.1). Nous vérifions ainsi l'hypothèse que certaines erreurs de génération apparaissent sous forme de séquences qui se répètent.

Nous avons ensuite analysé les prédictions correctes et incorrectes, leur adéquation par rapport au terme-source, ainsi que les relations lexicales entre ce dernier et sa reformulation, et la fonction sémantico-pragmatique de celle-ci. Pour ce faire, les 2266 prédictions générées automatiquement ont été annotées selon la grille présentée dans la section 4.2.

4.1 Traitement en n-grammes

Nous avons chargé les prédictions sur la plateforme SketchEngine, une plateforme proposant des outils avancés d'exploration de corpus. Nous avons appliqué une extraction de n-grammes à partir de prédictions générées automatiquement, dans l'objectif de vérifier les mots répétés et les syntagmes les plus fréquents dans les prédictions. Nous avons extrait les séquences de 2, 3, 4 et 5 n-grammes, avec un seuil de fréquence de 5. Parmi ces séquences, nous observons une préférence pour des prédictions contenant des termes génériques : *maladies, troubles, système*. Près d'un tiers des résultats (sur un total de 166 occurrences) a un terme générique comme tête lexicale.

Les plus longs (4 et 5 n-grammes) mettent en évidence des candidats qui sont en relation d'hyperonymie avec le terme : *les troubles mentaux chroniques, et/ou d'autres maladies, d'autres troubles*. Des syntagmes nominaux dont la tête est *maladie* ou *trouble* représentent 35 % des prédictions qui ont une longueur minimum de 4 mots. Parmi les résultats obtenus, on constate quelques répétitions du même mot (*maladie maladie de Parkinson*) en début de séquence. Cette anomalie est due à la présence de quelques séquences de ce type dans les données d'entraînement du système. L'outil de génération de texte a reproduit les répétitions repérées dans ces données d'entraînement.

L'extraction de n-grammes nous a permis d'observer la fréquence des candidats ainsi que le type de relation lexicale entre le terme-source et la reformulation. Ces éléments nous ont permis d'anticiper sur l'élaboration d'une grille d'annotation détaillée présentée ci-dessous.

4.2 Annotation des prédictions

Nous avons structuré les données sous la forme d'un tableur comprenant le terme-source, le terme-cible, c'est-à-dire, la production attendue, et la production effective, ou prédiction, par l'outil APT. Ces données brutes sont complétées par une première annotation évaluant le degré de correspondance de la prédiction APT vis-à-vis des attentes (Buhnila, 2023). Nous avons eu recours à une série d'annotations supplémentaires visant à décrire les prédictions générées automatiquement d'un point de vue formel, fonctionnel et sémantique. Nous nous sommes intéressées plus spécifiquement à la cohérence sémantique de la prédiction par rapport à la reformulation de la liste d'entraînement (*la vérité scientifique*) et aux incohérences des hallucinations générées automatiquement.

En effet, dans la littérature, les liens entre le terme-source et sa reformulation sont étudiés du point de vue de la relation lexicale (hyperonymie, hyponymie, synonymie, méronymie). Par exemple les reformulations paraphrastiques sont reliées par synonymie au terme-source (Vassiliadou, 2020), tandis que les définitions des termes médicaux contiennent des hyperonymes (Săpoiu, 2013 ; Bidu-Vrânceanu, 2007).

Les fonctions sémantico-pragmatiques représentent, quant à elles, les raisons qui poussent le locuteur à utiliser la reformulation. Eshkol-Taravella et Grabar (2017) identifient plusieurs fonctions sémantico-pragmatiques des reformulations : définition, paraphrase, exemplification, explication. Nous appliquons cette typologie pour identifier la fonction de chaque reformulation.

Nous avons conçu un guide d'annotation pour analyser les aberrations et les bonnes prédictions, annoter les reformulations adéquates, les relations lexicales et les fonctions sémantico-pragmatiques. Les hallucinations peuvent être abordées selon deux critères généraux : l'adéquation avec le terme (la prédiction générée peut être sémantico-référentiellement adaptée ou inadaptée) et la grammaticalité. Le critère de grammaticalité prend en compte plusieurs cas de figure : les mots inventés, les fautes d'accord, les constructions morphosyntaxiques erronées. La répétition du même terme est également relative à une

grammaticalité ténue (*maladie maladie de la peau*). Nous n'avons pas annoté les prédictions qui prennent la forme d'abréviations. Les annotations ont été réalisées par trois annotateurs linguistes.

Le critère d'adéquation permet de vérifier que la reformulation est en relation d'équivalence sémantique avec le terme-source. Si la prédiction est inadaptée, plusieurs raisons peuvent intervenir : i) la génération est correcte du point de vue grammatical mais il n'y a pas d'équivalence avec le terme (reformulation pour « insuffisance rénale » : *maladies de la peau*) ; ii) la génération est correcte du point de vue grammatical mais la combinatoire n'est pas adaptée ou contient des incohérences (*le stagiaire qui va se dissoudre rapidement*) ; iii) les mots inventés rendent également la prédiction inadaptée.

Tableau 2. Grille d'annotation des reformulations générées automatiquement.

Grammaticalité	**Adéqua-tion**	**Type inadéquation**	**Fonction sémantico-pragmatique**	**Relations lexicales**	**Marqueurs explicitement présents**	**Patrons morpho-syntaxiques**
Correct **Incorrect** **Partiellement correct** ***(tête de syntagme correct, ex. maladie d'allure schizophrénique)***	Oui/Non	Non-F *(formellement non conforme mais compréhensible)* Non-E *(correct mais non équivalent)* Non-S *(sémantiquement inacceptable)*	Catégorisation Comparaison Définition Exemplification Explication Paraphrase	Hyperonymie / Hyponymie Synonymie Méronymie	Ex. *tel que, comme, c'est-à-dire…*	Ex. N+Adj GN Etc.

Les annotations nous permettent d'identifier les inadéquations sémantico-référentielles, généralement assimilées à des hallucinations, et de vérifier plusieurs hypothèses concernant i) le rôle de la relation hyperonymique et de la généralisation ; ii) les types de patrons mobilisés, en fonction des marqueurs explicitement présents, ainsi que iii) le type de fonction sémantico-pragmatique le plus souvent associée à une prédiction de type « hallucination ».

5 Analyse quantitative et qualitative des reformulations générées automatiquement

Nous avons conduit une analyse de l'ensemble des prédictions, avec un focus spécifique sur les inadéquations sémantiques. L'analyse quantitative et qualitative détaillée dans les sections qui suivent nous permettra d'évaluer la prégnance des patrons appris de manière statistique par l'algorithme et, plus spécifiquement, les patrons les plus sensibles aux inadéquations sémantiques.

5.1 Analyse quantitative

5.1.1 Degré d'adéquation des prédictions

Une première analyse évalue le pourcentage des prédictions qui sont correctes du point de vue grammatical. Nous avons exploité trois catégories d'étiquettes : la prédiction est correcte (parfaitement grammaticale), la prédiction est incorrecte (la forme est agrammaticale : manque les déterminants, l'accord entre sujet et prédicat n'est pas réalisé), la prédiction est partiellement correcte (ex. : mot inventé, mais qui s'inscrit dans la phrase et respecte les règles de formation du mot). Nous avons 2266 prédictions au total, pour 480 termes-sources.

Tableau 3. Répartition des prédictions grammaticalement correctes, incorrectes, partiellement correctes.

Ensemble des prédictions	Prédictions correctes	Prédictions incorrectes	Abréviation	Répétition	Partiellement correctes	Sans étiquette
2266 (100%)	725 (31,99 %)	779 (34,38 %)	1 (0,05 %)	5 (0,22 %)	211 (9,31 %)	545 (24,05 %)

Sur 2266 prédictions annotées, 31,99 % sont correctes du point de vue grammatical. 34,38 % sont des prédictions incorrectes et 9,31 % sont partiellement correctes (présence de mots inventés). 24,05 % ont été exclus de l'analyse, car la prédiction était une abréviation. Les prédictions grammaticalement incorrectes sont majoritaires ; les fautes d'accord et l'absence de déterminants sont les erreurs les plus fréquentes.

Un deuxième critère d'analyse concerne la capacité à produire une reformulation adaptée. Une grande majorité de prédictions est inadaptée, soit 63,56 % pour seulement 11,33 % de prédictions adaptées. Parmi les prédictions qui sont inadaptées, on constate l'existence de plusieurs cas : la prédiction proposée n'est pas l'équivalent sémantique du terme (38,28 %) ; 24,48 % des prédictions contiennent des combinaisons impossibles du point de vue sémantique ; ou encore, dans 1,58 % de cas, la reformulation n'est pas tout à fait adaptée, mais il est possible de comprendre le sens. Certaines prédictions, en plus des abréviations, n'ont pas pu être catégorisées : il s'agit de répétitions à l'identique ou de plusieurs répétitions de mots qui sont inclus dans le terme original.

5.1.2 Fonctions sémantico-pragmatiques et relations lexicales

Le **Tableau 4** résume les fréquences des fonctions sémantico-pragmatiques et des relations lexicales dans l'ensemble des prédictions. Nous remarquons que les fonctions les plus présentes dans les prédictions sont les exemplifications et les définitions, ce qui est confirmé également par les relations lexicales correspondantes, tels que l'hyponymie et l'hypéronymie (Buhnila, 2022). Les explications sont les moins fréquentes parmi les reformulations générées, de même que la relation de méronymie, qui permet d'expliquer un terme ou une procédure médicale par des parties du concept global.

Tableau 4. Analyse quantitative des fonctions sémantico-pragmatiques (exemplification, définition, catégorisation, paraphrase, explication) et des relations lexicales (hyperonymie, hyponymie, synonymie, méronymie) des prédictions de reformulations correctes et incorrectes. *1,99 % et **3,17 % des prédictions incorrectes n'ont pu être attribuées aucune relation ou fonction par les annotateurs à cause de leurs formes grammaticalement incorrectes.

	Fonctions sémantico-pragmatiques					**Relations lexicales**			
	exempl.	**déf.**	**catég.**	**paraphr.**	**explic.**	**hyperony.**	**hypony.**	**syn.**	**mérony.**
N°	325	255	214	77	76	660	189	93	13
%	33,64	26,39	22,15	7,97	7,86	66,93	19,16	9,43	1,31
Ensemble des prédictions		966 (100%)*					986 (100%)**		

Le **Tableau 5** illustre la distribution de chaque fonction et relation par rapport au type de prédiction, correcte ou incorrecte. En comparant les prédictions correctes avec les prédictions incorrectes, nous observons que la fonction la plus fréquente dans les deux catégories est l'exemplification, avec une distribution presque équivalente (34,33 % et 33,15 %). Cette fonction est suivie par la définition, pour laquelle nous observons un nombre plus grand de prédictions incorrectes (27,16 % incorrectes contre 25,31 % correctes). Les paraphrases et les explications incorrectes sont également plus fréquentes, ce qui peut s'expliquer par la difficulté de générer automatiquement ce type de reformulations qui ont des formes très variables.

Tableau 5. Distribution des fonctions sémantico-pragmatiques et des relations lexicales parmi les prédictions correctes et les « hallucinations » (selon les annotations des linguistes).

	Prédictions correctes[v]		**Prédictions incorrectes**		**Ensemble des prédictions**[vi]	
Fonctions sémantico-pragmatiques						
exemplification	137	34,33%	188	33,15%	325	33,64%
définition	101	25,31%	154	27,16%	255	26,39%
catégorisation	90	22,55%	124	21,86%	214	22,15%
explication	35	8,77%	41	7,23%	76	7,86%
paraphrase	22	5,51%	55	9,70%	77	7,97%
TOTAL	**399**	**41,30%**	**567**	**58,69%**	**966**	**100%**
Relations lexicales						
hyperonymie	344	80,56%	332	59,39%	660	66,93%
hyponymie	38	8,89%	151	27,01%	189	19,16%
synonymie	38	8,89%	55	9,83%	93	9,43%
méronymie	0	0%	13	2,32%	13	1,31%
TOTAL	**427**	**43,30%**	**559**	**56,69%**	**986**	**100%**

Nous remarquons la tendance de l'outil à surutiliser la relation d'hyperonymie dans les prédictions correctes pour jusqu'à 80,56 % des reformulations générées automatiquement. L'hyperonymie est également la plus fréquente dans les générations inadéquates (59,39 %), ce qui montre que l'outil a surappris les patrons des reformulations introduites par des hyperonymes (avec des marqueurs de type *est une maladie, défini comme une maladie*). La relation d'hyponymie est surreprésentée dans les « hallucinations » (27,01 %, contre 8,89 % pour les prédictions correctes). L'outil reproduit le patron de l'exemplification par des hyponymes (exemple 1 ci-dessous) mais, dans la grande majorité des cas, ces prédictions sont incorrectes (76,59 % prédictions de type exemplification avec hyponymie). La majorité des incohérences sont d'ordre grammatical ou formel (82,63 %), tandis que 17,36 % sont des incohérences sémantiques (exemple 2).

(1) Terme : troubles du cou
Reformulation originale : tels que les troubles associés au coup de fouet cervical
Prédiction : tels que la thrombose cérébrale, le trouble cardiaque chronique et les troubles musculo-squelettiques

(2) Terme : strabisme
Reformulation originale : est un/e affection dans laquelle les yeux ne sont pas alignés normalement
Prédiction : est un/e maladie associée à une stagiaire ayant tendance à se dissoudre rapidement

Nous comparons ces résultats avec les annotations des relations lexicales et fonctions sémantico-pragmatiques du jeu de données d'entraînement et le jeu de données de test du modèle (**Figures 1** et **2**). Nous observons que le modèle de langue génère des structures formelles proches de celles apprises lors de l'étape d'entraînement. Les paires de relation-fonction *hyperonymie-définition* et *hyponymie-exemplification* sont les plus fréquentes dans les données d'entraînement, tandis que dans les données de test la paire *hyperonymie-paraphrase* (ex. maladie musculaire progressive = troubles de déglutition) est plus représentée que celle d'*hyponymie-exemplification*.

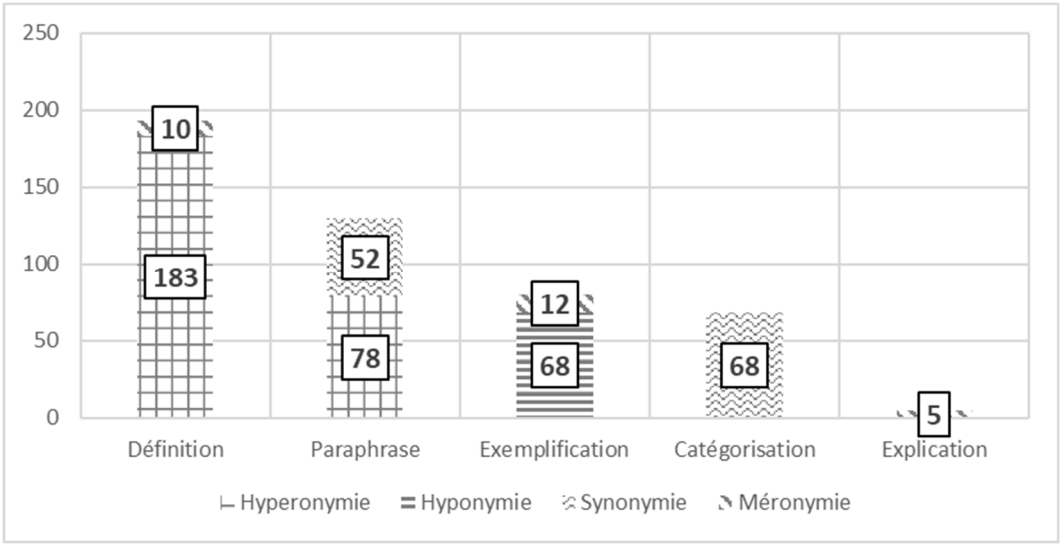

Figure 1. Relations lexicales et fonctions sémantico-pragmatiques du jeu de données de test (liste de 480 paires de termes médicaux avec leurs reformulations correctes correspondantes).

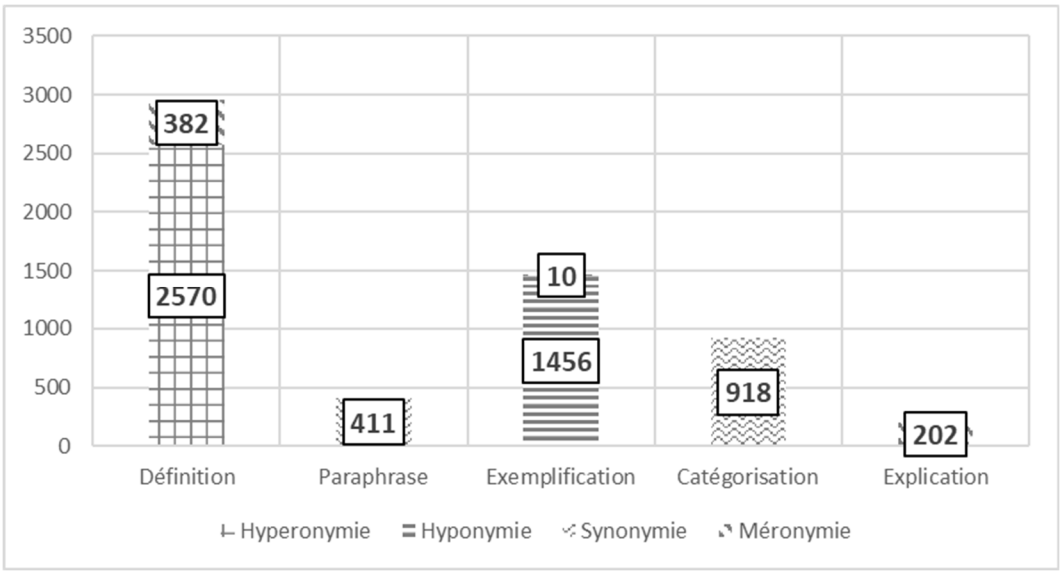

Figure 2. Relations lexicales et fonctions sémantico-pragmatiques du jeu de données d'entraînement. L'annotation présentée a été validée par un accord inter-annotateur Kappa (Cohen, 1960) entre deux annotateurs linguistes francophones sur 5945 paires de termes médicaux - reformulations correctes en français.

5.1.3 Exploitation des patrons morphosyntaxiques

Enfin, l'annotation des patrons syntaxiques a été réalisée sur 1261 prédictions, dont 578 (45,83 %) correctes et 683 (54,16 %) incorrectes. Les patrons syntaxiques les plus fréquents générés par l'algorithme pour l'ensemble des annotations sont basés sur des structures nominales, de type groupe nominal (GN) suivi par un adjectif (12,68 % ; *c'est-à-dire un événement indésirable*), GN suivi par un groupe prépositionnel (GP) (10,07 % ; *défini comme une augmentation de 2 à 10 %*) et des suites d'au moins deux adjectifs (9,05 % ; *thrombose thrombotique allo-immunisée*) (**Tableau 6**). Parfois, l'outil emploie des noms sans déterminants suivis par des adjectifs ou des GP, de type *(douleurs aigus de cette affection)* (11,57 %). Cette tendance peut être due aux données d'entraînement, qui comportent le format sans déterminant. Les verbes sont très peu présents dans les reformulations. Nous avons remarqué une forte utilisation des conjonctions de coordination de type *et*, *ou* dans les prédictions générées (26,32 % ; *les troubles du comportement et la pression artérielle ; d'oreillons rouges, rouges rouges ou rouges bleus*).

Tableau 6. Données quantitatives de l'analyse des patrons syntaxiques des prédictions générées automatiquement. N* représente les noms sans déterminants générés par le modèle de langue.

Patrons syntaxiques	N° de prédictions	Pourcentage
GN + GP/GN/ADJ + CONJ (et/ou)	332	26,32 %
GN + ADJ	160	12,68 %
N* + ADJ/GP	146	11,57 %
GN + GP	127	10,07 %
GN + ADJ + ADJ	114	9,05 %

5.2 Entrée par les relations lexicales

Les annotations des relations lexicales concernent 67,29 % des prédictions réalisées automatiquement par le système. Parmi les annotations de ces relations, 29,78 % sont des relations d'hyperonymie, 8,34 % des relations d'hyponymie et seulement 4,10 % des relations de synonymie, 0,66 % des relations de méronymie. Un pourcentage important (24,40 %) représente des cas où il est difficile d'identifier une relation lexicale, soit parce que des connaissances expertes du domaine médical sont nécessaires, soit en raison de la présence de mots inventés.

Sur 675 prédictions annotées en hyperonymie, seulement 34,66 % sont grammaticalement correctes. La plupart des hyperonymes proposés ont des termes génériques comme têtes lexicales (*maladie, trouble, affection, médicament, traitement*). Parmi les prédictions correctes du point de vue grammatical, mais qui sont inadaptées, on retrouve les mêmes termes génériques accompagnés d'un adjectif ou d'un complément de nom, mais dont la combinaison n'est pas adaptée en tant que reformulation, comme dans l'exemple suivant :

(3) Terme : myosite à inclusions
Reformulation correcte : , maladie musculaire
Prédiction : maladies des voies musculaires intraveineuses.

Dans l'exemple 4, c'est le terme proposé comme hyperonyme qui n'est pas adapté (*coagulopathie vs virus*).

(4) Terme : coagulopathie
Reformulation correcte : (trouble hémostatique)
Prédiction : le virus de l'hémoglobine

16,93 % de prédictions annotées en hyponymie sont correctes (sur les 175 présentes dans le corpus). Les prédictions incorrectes représentent 83,07 % des cas. Le système propose des exemples plus spécifiques que le terme (exemple 5), soit des noms de maladies inventés ou des combinaisons lexicales étranges, en plus de l'utilisation des mots commençant par les mêmes syllabes « myo- » (exemple 6) :

(5) Terme : troubles gastro-intestinaux

Reformulation correcte : , effets secondaires
Prédiction : comme la douleur gastro-intestinale

(6) Terme : myopie
Reformulation correcte :est un/e défaut de la vision qui se trouble lorsque des objets sont observés à distance
Prédiction :, tel que un sombre myocarde

Les relations de synonymie sont plus rarement identifiées, mais souvent associées avec la fonction de paraphrase et sont correctes dans 33,33 % des cas. Le terme synonyme peut être introduit par un terme générique (« maladie delirium » synonyme de « syndrome confusionnel ») ou alors par une séquence explicative (« modification de la composition de l'alimentation » synonyme de « changement de l'alimentation »).

Les résultats de l'analyse quantitative concordent avec les observations spécifiques pour chaque type de relations. Dans le cas des hyperonymes et des synonymes, on utilise des termes génériques (maladie, trouble, affection, syndrome etc.) pour proposer des reformulations. Le pourcentage de prédictions correctes est similaire dans ces deux cas. Le système semble être plus performant au niveau des généralisations que pour proposer des exemples plus spécifiques (dans le cas des hyponymes).

5.3 Entrée par les fonctions sémantico-pragmatiques

5.3.1 Exemplification et illustration

Nous avons fait l'hypothèse que la fonction d'exemplification est plus souvent associée à la relation d'hyponymie, d'après les constats présentés dans la littérature. Dans nos données (187 exemplifications), la relation d'hyponymie est en effet davantage exploitée que celle d'hyperonymie (18,47 % contre 5,80 %). Elle est très souvent associée à une énumération de GN ou de plusieurs N, ou encore introduite par le marqueur *par exemple, tel que.* Le double marqueur *par exemple* + parenthèses *()* est associé à l'exemplification et à l'hyponymie dans seulement 3,06 % des cas. Les hyponymes proposés peuvent représenter des noms spécifiques de maladies, de symptômes ou de traitement.

Souvent, on retrouve des exemplifications dans les prédictions correctes et adaptées :

(7) Terme : maladies sexuellement transmissibles
Prédiction : , par exemple le VIH/SIDA

Parfois, la prédiction est incorrecte et/ou inadaptée :

(8) Terme : maladie de décompression
Prédiction : , tel que l'endométriose

Les prédictions sont inadaptées pour non-équivalence, mais aussi à cause des combinaisons sémantiques hasardeuses. La première catégorie est massivement associée avec l'exemplification et l'hyponymie. La non-équivalence est en cause dans l'exemple suivant, la reformulation est une série d'hyponymes mais qui ne sont pas complètement équivalents :

(9) Terme : états pathologiques
Prédiction : (tel que des maladies cérébrales et un déséquilibre pathologique, des troubles musculosquelettiques ou lésions cérébrales sévères)

Les combinaisons sémantiques inadaptées sont parfois dues aux mots inventés :

(10) Terme : traitement par suppression androgénique
Prédiction : (par exemple l'anorexination)

5.3.2 Définitions et contextes définitoires

Nous analysons les hallucinations de type définition en partant des marqueurs de reformulations de type *défini*, *défini comme*, *définition*, *est un/e* et *c'est-à-dire*. Nous analysons la fréquence et la structure des aberrations afin d'identifier des possibles patrons dans la génération automatique de ces résultats.

Tableau 7. Fréquence des marqueurs dans les prédictions de reformulations générés automatiquement.

Marqueur	Ensemble des prédictions	Prédictions correctes	Prédictions incorrectes	Répétition du terme
défini/e ; défini/e comme	184	18	165 (89,67 %)	2
est un/e ; sont des/une	66	6	60 (90,90 %)	0
c'est-à-dire	122	7	106 (86,88 %)	9
en particulier	22	4	16 (72,72 %)	1

Nous remarquons une fréquence élevée des générations introduites par le lemme *défini* : 184 prédictions, dont 161 prédictions marquées par *défini/e comme* (exemple 12 et 14). En moyenne, 85 % des prédictions générées pour les marqueurs étudiés sont des aberrations.

Tableau 8. Exemples des prédictions de reformulations correctes (1) et des prédictions hallucinatoires (0).

Terme médical	Reformulation correcte	Prédiction générée automatiquement	Statut	Fonction sémantico-pragmatique
déclin cognitif	ou démence	(11) (définition désigne les états cognitifs les plus répandus)	0	
		(12) (défini/e ; définie comme atteinte d'une mentalité déficiente et d'une discrimination sur le plan cognitif ou émotionnel)	1	définition
cardiopathie ou un diabète	maladie chronique	(13) c'est-à-dire des troubles cardiaques ou diabétiques	1	exemplification
critères de jugement cliniques réels	les maladies micro et macrovasculaires	(14) (défini/e comme les résultats obtenus après un exercice et la cote de référence)	0	
		(15) (c'est-à-dire la nécessité de régler les problèmes de santé en se basant sur des études, des tests et des études de référence)	1	définition
acétylcholine	connue sous le nom de substance chimique	(16) c'est-à-dire l'oxyde d'oxalate de glycosyle de l'oxalate	0	
kératoses actiniques	est un/e maladie de peau causée par l'exposition au soleil sur le long terme	(17) c'est-à-dire des lésions de l'appareil pénible	0	
médecine traditionnelle chinoise	(MTC)	(18) (MTC)	2	abréviation
		(19) (c'est-à-dire les méthodes d'admission orales dans les écoles ou les foyers de l'école, par exemple)	0	
		(20) (c'est-à-dire une ancienne méthode chinoise)	1	paraphrase

5.3.3 Paraphrases, explications

Le marqueur *c'est-à-dire* peut avoir d'autres fonctions sémantico-pragmatiques en plus de la définition (exemple 15), dont celle de *la paraphrase synonymique* ou de *l'explication*. Si *c'est-à-dire* est un marqueur typique de la reformulation (Fuchs, 1982 ; Vassiliadou, 2020), il n'est pas très fréquemment utilisé par l'outil. APT utilise ce marqueur pour un total de 71 prédictions, dont 50 explications, 7 paraphrases synonymiques, 6 définitions / hypéronymie, 5 exemplifications. Les exemples 16, 17 et 19 montrent des

prédictions hallucinatoires, avec des mots inventés (*l'oxyde d'oxalate*) ou sémantiquement incohérentes (*lésions de l'appareil pénible*).

5.3.4 Relation de comparaison/catégorisation

L'outil exploite des marqueurs servant à mettre en série des éléments, à l'instar de *et d'autres*, produisant ainsi, au-delà de l'effet de liste, un effet de catégorisation. L'utilisation d'adjectifs relationnels (*rhumatismales, schizophrénique*) peut également produire un effet de catégorisation. Ces stratégies génératives vont de pair avec l'exploitation régulière des relations d'hyperonymie. Sur l'ensemble, on recense 214 occurrences, soit 9,4% de ce type de relation dans le corpus.

5.4 Analyse qualitative : où l'on apprend qu'une *toux non conforme aux règles d'élevage sévère* est une *toux rouge*

Peut-on parler d'inadéquation sémantique dans le cas des aberrations générées automatiquement ? Les prédictions de l'outil étant le résultat d'une identification de patrons et de régularités combinatoires dans le corpus d'apprentissage et leur reproduction, le but de cette exploration qualitative est de saisir les régularités ayant potentiellement induit ces aberrations. Ainsi, par exemple, dans le cas f) ci-dessous, « hématologie », qui est le nom d'un domaine de spécialité, et « inflammatoire », qui accompagne généralement des noms d'affections, constituent une association fautive. Cependant, une recherche sur le Web montre des cas de voisinage du type « […] hématologie, inflammatoire […] » qui peuvent expliquer l'occurrence produite par l'outil.

La vraie question serait : peut-on parler de sémantique ? S'agissant de générations artificielles de séquences langagières exploitant la combinatoire de patrons langagiers, la notion de « sens », qui présuppose l'intention de faire sens, n'est probablement pas adaptée. Mais ce sont justement les conflits combinatoires qui produisent ce qui est interprété comme une inadéquation sémantique et donne lieu au terme – discutable – d'hallucination.

Nous avons recensé plusieurs types de conflits combinatoires :

a) L'utilisation de termes venant de l'anglais, ex. *défini as le retard ; maladie, affection du skin.*
b) L'invention de termes exploitant des morphèmes ou lexèmes du corpus d'apprentissage, ex. *l'arthrite bulléenne acquise et la bêta-pythéliale ; , le dexamethasone, le périglycone, le pénétrolythétride, le lécithylbenzène, le d-glucone, le bromoquinoline, le boscitol et le terpenol.*
c) L'utilisation de termes ne présentant aucune affinité référentielle avec le domaine recherché, ex. *les organismes d'envergure mondiale*.
d) La répétition de termes, ex. *définie comme une complication inflammatoire et répétitive de la peau ou de la peau ou de la peau acquise d'une personne et qui touche une personne ou un groupe de personnes qui ne sont pas aptes à faire des choix alimentaires ou d'autres activités.*
e) L'association de termes à sèmes incompatibles, ex. *toux rouge (qui n'est pas sans rappeler « d'incolores idées vertes dorment furieusement » de Chomsky) ; (par exemple une fente dégradante de la santé) ; (par exemple les exercices de physiothérapie ou la guitare) ; (par exemple thérapie vocale de la colonne vertébrale du nez et de la tête).*
f) L'association de termes dont l'évaluation de la compatibilité nécessite une connaissance terminologique, ex., *des hématologies inflammatoires et vasculaires et des infections à germes rouges*.

L'analyse combinatoire est compliquée par le fait que certaines productions relèvent de cas de figures mixtes, comme l'exemple ci-dessous, qui emprunte à c) et e) et où l'inadéquation réside moins dans l'incompatibilité à l'intérieur d'un groupe syntaxique et davantage dans l'incompatibilité relationnelle entre les groupes connectés par *et/ou* et leur non-informativité :

(21) (par exemple les interventions sur des maladies non affectives et des interventions sur des maladies mentales, notamment des soins obstétriques ou de la médecine), ou il n'y a pas d'enquête suffisante sur les troubles musculosquelettiques, d'enquête sur l'exposition à des sédations ou sur des autres questions liées aux symptômes ou aux troubles sociaux d'un enfant à l'âge adult

La notion de *lexical priming* nous semble la plus adaptée pour aborder les conflits sémiques :

> As a word is acquired through encounters with it in speech or writing, it becomes cumulatively loaded with the contexts and co-texts in which it is encountered, and our knowledge of it includes the fact that it co-occurs with certain other word in certain kinds of context. The same applies to word sequences built out of these words; these too become loaded with the contexts and co-texts in which they occur. (Hoey, 2005 : 8)[vii]

Telle que définie, la notion prévoit des attentes en termes de combinatoire en mettant l'accent sur le contenu sémantique véhiculé. À la différence d'autres notions proches comme celle de prosodie sémantique (Sinclair, 1996) ou d'isotopie sémantique (Rastier, 1985), la notion de *lexical priming* ne met pas en exergue le positionnement du locuteur – impossible, compte tenu de nos données, ni l'harmonie sémantique – discutable également. Ainsi, dans *col de la peau*, « col » vient avec une série de contextes d'usage prédéfinis, montrant une préférence collocationnelle pour des termes comme « montagne » ou « utérus », en fonction du domaine d'usage, et invalidant la séquence générée. Inversement, la séquence *dématérialisation en phase terminale* pourrait presque bénéficier d'une lecture poétique, compte tenu des co-textes habituellement associés à « phase terminale » et du trait [disparition] potentiellement partagé entre « dématérialisation » et « mort ».

On constate par ailleurs, dans certaines prédictions, un alignement de connotations négatives, moyennant l'usage de termes à polarité négative qui restent, pour la plupart, non spécialisés, produisant une impression de prosodie sémantique :

(22) (défini/e ; définie comme une affection non néphrologique, une augmentation des taux d'injection, la présence d'insuffisance rénale et l'augmentation des taux de diarrhée, un désavantage génétique, un retard et une faible pression sanguine

Ces différents exemples soulèvent au moins trois questions. La première est celle du sens, de son rapport à la fois à l'intentionnalité et à l'interprétation en réception. Lorsque les séquences générées sont hors-sujet ou les erreurs combinatoires grossières, il est plus facile de constater d'emblée l'inadéquation sémantique. La deuxième question est celle des ressources mobilisables. Ainsi, lorsque l'interprétation nécessite des connaissances de spécialité, l'évaluation est beaucoup plus difficile. Enfin, la troisième question concerne les principes et les frontières de la créativité. Lorsque la combinatoire s'écarte de peu par rapport à ce qui est attendu, certaines séquences générées automatiquement peuvent produire un effet stylistique.

Nos observations seraient à approfondir sur la base d'annotations sémantiques afin de pouvoir dégager des tendances générales. Se pose également la question d'approches automatisées de l'analyse sémantique.

6 Conclusion et perspectives

Le taux de reformulations soit grammaticalement incorrectes soit sémantico-référentiellement inadéquates est élevé dans notre corpus (moins de 12% de reformulations adaptées). Compte tenu du domaine spécialisé dans le champ médical, l'évaluation de l'adéquation de la combinatoire lexicale et donc des reformulations nécessite dans un quart des cas une analyse et une interprétation fines et documentées. Grâce aux annotations, nous avons pu constater que dans sa production de reformulations, l'outil marque une préférence statistique pour l'usage des hyperonymes, présents dans 80,56 % des reformulations correctes et 59,39% des reformulations inadéquates. L'exemplification favorise quant à elle la surutilisation des hyperonymes. L'annotation morphosyntaxique d'un échantillon a mis en évidence l'utilisation de conjonctions de coordination du type *et/ou* dans 26,32% des cas – *et* introduit régulièrement des répétitions ou des combinaisons inadéquates.

L'habileté de généralisation des modèles LLM est généralement reconnue et exploitée dans le domaine des générateurs textuels (Yang et al. 2024). Dans les exemples d'aberrations analysés, nous avons pu constater des proximités avec l'exploitation des patterns dans quelques situations de pratiques langagières naturelles. Ainsi, les outils de génération textuelle peuvent avoir recours à des généralisations abusives en mettant en exergue des patrons dont la portée n'est pas validée par l'usage. Des processus comparables sont identifiés :

- chez les apprenants d'une langue étrangère ;
- dans certains calculs de fréquence (cf. Piantadosi (2014) sur la loi de Zipf).

Dans les deux cas, l'identification de quelques patterns récurrents peut produire une généralisation abusive : au niveau de l'usage dans le cas des apprenants, au niveau de l'interprétation dans le cas de l'exploration des données à l'aide d'outils informatiques (et en l'absence de pondérations statistiques contextualisées). Chez l'humain, la répétition est connue pour produire un effet de vérité (Hasher et al., 1977 ; Dechêne et al., 2010) qui prend la forme d'un abus de langage chez la machine.

Dans les données analysées, les cas où cela se rapproche le plus de la réalité des discours, ou encore lorsque l'interprétation nécessite la maîtrise de connaissances médicales, soulèvent le plus de questionnements, dans la mesure où ils peuvent produire un effet de « vérité » - qui n'est rien d'autre qu'un effet rhétorique, le seul objectivement accessible aux outils de génération langagière (cf. Bender et Koller, 2020).

Si d'un point de vue informatique les objets que nous venons d'examiner soulèvent des questions quant au paramétrage des modèles langagiers exploités par les réseaux neuronaux, d'un point de vue linguistique nos observations soulèvent la question de l'interface entre les usages purement linguistiques, qui donnent lieu à des patterns identifiables, et leurs différents contextes sur le terrain des pratiques langagières qui ancrent ces patterns dans l'usage. Cela ne peut que rejoindre les recommandations de Bender et al. (2021) d'utiliser des données propres et soigneusement sélectionnée pour l'entraînement des outils génératifs. Dans cette même optique, les données utilisées pour nos expériences seront partagées avec la communauté afin de permettre la reproduction de nos résultats et l'amélioration des modèles de langues sur la tâche de génération de paraphrases sous-phrastiques médicales en français. (https://github.com/ibuhnila/refomed)

Références bibliographiques

Alkaissi, H. et McFarlane, SI. (2023). Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus* 15(2): e35179. DOI: 10.7759/cureus.35179

Buhnila Ioana. (2023). *Une méthode automatique de construction de corpus de reformulation*. Thèse de doctorat, Université de Strasbourg, juin 2023.

Athaluri, SA., Manthena, SV., Kesapragada VSR, KM., Yarlagadda, V., Tirth, D. et Rama, TSD. (2023). Exploring the Boundaries of Reality: Investigating the Phenomenon of Artificial Intelligence Hallucination in Scientific Writing Through ChatGPT References. *Cureus* 15(4): e37432. DOI 10.7759/cureus.37432

Bender, EM., Gebru, T., McMillan-Major, A. et Shmargaret, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, p. 610-623. https://doi.org/10.1145/3442188.3445922

Bender, E., Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 5185-5198.

Bidu-Vrânceanu, A. (2007). *Lexicul specializat în mișcare. De la dicționare la texte*. București. *Editura Universității din București*. 266 pages.

Bruno, A., Mazzeo PL., Chetouani, A., Tliba, M., Kerkouri, MA. (2023). Insights into Classifying and Mitigating LLMs' Hallucinations. *arXiv*:2311.08117v1 [cs.CL]

Buhnila, I. (2022). Le rôle des marqueurs et indicateurs dans l'analyse lexicale et sémantico-pragmatique de reformulations médicales. *8e Congrès Mondial de Linguistique Française (CMLF)*, 4-8 juillet 2022, Orléans, France, SHS Web of Conferences 138: 10005. https://doi.org/10.1051/shsconf/202213810005.

Bybee, J. (2006). From Usage to Grammar: The Mind's Response to Repetition. *Language, 82(4)*, p. 711-733.

Chomsky, N. (1957). *Syntactic Structure*. Mouton.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educ. Psychol. Meas*., 20, p. 27-46.

Copara, J., Knafou, J., Naderi, N., Moro, C., Ruch, P. et Teodoro, D. (2020). Contextualized French language models for biomedical named entity recognition. *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes*, p. 36-48.

Culbertson, J., Schouwstra, M. et Kirby, S. (2020). From the world to word order: Deriving biases in noun phrase order from statistical properties of the world. *Language* 96(3), p. 1-22.

De Castro, M., Zona, U. (2022). A vigotskijan perspective on machine learning. How cultural stereotypes are involved in education of algorithms. *Academia Letters*, Article 4638. https://doi.org/10.20935/AL4638.

Dechêne, A., Stahl, C., Hansen, J. et Wänke, M. (2010). The truth about the truth: A meta-analytic review of the truth effect. *Personality and Social Psychology Review* 14(2), p. 238-257. doi:10.1177/1088868309352251.

Devlin, J., Chang, M-W., Lee, K. et Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Durt, C., Froese, T., Fuchs, T. (2023). Against AI Understanding and Sentience: Large Language Models, Meaning, and the Patterns of Human Language Use. [Preprint] *PhilSci Archive*.

Eddine, M. K., Tixier, A., Vazirgiannis, M. (2021). BARThez: a Skilled Pretrained French Sequence-to-Sequence Model. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing,* p. 9369-9390.

Emsley, R. (2023). ChatGPT: these are not hallucinations – they're fabrications and falsifications. *Schizophrenia* 9:52. https://doi.org/10.1038/s41537-023-00379-4.

Eshkol-Taravella, I., Grabar, N. (2017). Taxinomie dans les reformulations du point de vue de la linguistique de corpus. *Syntaxe et Sémantique,* vol. 18, no. 1, p. 149-184.

Fuchs, C. (1982). La paraphrase entre la langue et le discours. *Langue française, La vulgarisation* (53), p. 22-33.

Goldberg, A. (2019). *Explain Me This: Creativity, competition, and the partial productivity of constructions*. Princeton University Press.

Grabar, N., Cardon, R. (2018). CLEAR – Simple Corpus for Medical French. *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, Tilburg, the Netherlands. Association for Computational Linguistics, p. 3-9.

Gülich, E., Kotschi, T. (1983). Les marqueurs de la reformulation paraphrastique. *Cahiers de linguistique française* 5, p. 305-351.

Hasher, L., Goldstein, D. , Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior* 16(1), p. 107-112. doi:10.1016/S0022-5371(77)80012-1

Hatem R., Simmons B., Thornton JE. (2023). Chatbot Confabulations Are Not Hallucinations. *JAMA Intern Med.* 2023, 183(10):1177. doi:10.1001/jamainternmed.2023.4231

Heidegger, M. (2010). Being and Time. Translated by Joan Stambaugh and Dennis J. Schmidt. *SUNY Series in Contemporary Continental Philosophy.* Albany: State University of New York Press.

Hoey, M. (2005). *Lexical Priming: A new theory of words and language*. Abingdon, England: Routledge.

Hopper, P., Bybee, J. (2001). *Frequency and the Emergence of Linguistic Structure*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. et Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography 1*, p. 7-36.

Labrak, Y., Bazoge, A., Dufour, R., Rouvier, M., Morin, E., Daille, B. et Gourraud, P. A. (2023). DrBERT: Un modèle robuste pré-entraîné en français pour les domaines biomédical et clinique. *18e Conférence en Recherche d'Information et Applications\\16e Rencontres Jeunes Chercheurs en RI\\30e Conférence sur le Traitement Automatique des Langues Naturelles\\25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, p. 109-120.

Lin, C. Y. (2004). Rouge: A package for automatic evaluation of summaries. *Text summarization branches out*, p. 74-81.

Martin, L., Muller, B., Ortiz Suárez P.J, Dupont, Y., Romary, L., de la Clergerie, E., Seddah, D., Sagot, B. (2020). CamemBERT: a Tasty French Language Model. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics, p. 7203-7219.

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J. et Kiela, D. (2020). Adversarial nli: A new benchmark for natural language understanding (https://arxiv.org/abs/1910.14599).

Nighojkar, A., Licato, J. (2021). *Improving paraphrase detection with the adversarial paraphrasing task*. arXiv preprint. arXiv:2106.07691.

Østergaard, SD., Nielbo, KL. (2023). False Responses From Artificial Intelligence Models Are Not Hallucinations. *Schizophrenia Bulletin,* Volume 49, Issue 5, p. 1105-1107, https://doi.org/10.1093/schbul/sbad068

Palivela, H. (2021). Optimization of paraphrase generation and identification using language models in natural language processing. *International Journal of Information Management Data Insights*, 1(2), 100025.

Piantadosi, ST. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin et Review* 2, p. 1112–1130. https://doi.org/10.3758/s13423-014-0585-6

Post, M. (2018). A Call for Clarity in Reporting BLEU Scores. *Proceedings of the Third Conference on Machine Translation: Research Papers*, p. 186-191.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S, Matena, M., Zhou, Y., Li, W. et Liu, PJ. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1), p. 5485-5551.

Rastier, F. (1985). *L'isotopie sémantique, du mot au texte*. Paris.

Săpoiu, C. (2013*). Hiponimia în terminologia medicală. Modalităţi de abordare în semantică şi lexicografie*. Piteşti, Editura Trend, 199 pages.

Sellam, T., Das, D. et Parikh, AP. (2020). Bleurt: Learning robust metrics for text generation. arXiv preprint arXiv:2004.04696.

Sinclair, J. (1996). The search for units of meaning. *Textus* 9, p. 75–106.

Tchechmedjiev, A., Abdaoui, A., Emonet, V., Zevio S. et Jonquet, C. (2018). SIFR annotator: ontology-based semantic annotation of French biomedical text and clinical notes. *BMC bioinformatics, 19(1)*, 405.

Todirascu, A., Padó, S., Krisch, J., Kisselew, M. et Heid, U. (2012). French and german corpora for audience-based text type classification. *LREC, volume 2012*, p. 1591-1597.

Touchent, R., Romary, L. et De La Clergerie, E. (2023). CamemBERT-bio: Un modèle de langue français savoureux et meilleur pour la santé. *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1: travaux de recherche originaux--articles longs*, p. 323-334.

Vassiliadou, H. (2020). Peut-on aborder la notion de "reformulation" autrement que par la typologie des marqueurs? pour une analyse sémasiologique et onomasiologique. In Olga Inkova (ed.), *Autour de la Reformulation,* Droz, p. 77-94.

Vernikos, G., Popescu-Belis, A. (2024). Don't Rank, Combine! Combining Machine Translation Hypotheses Using Quality Estimation. *arXiv preprint arXiv:2401.06688*.

Witteveen, S., AI, R. D., Andrews, M. (2019). Paraphrasing with Large Language Models. In *Proceedings of the 3rd Workshop on Neural Generation and Translation, EMNLP-IJCNLP 2019*, p. 215-220.

Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Zhong, S., Yin, B. et Hu, X. (2024). Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. *ACM Trans. Knowl. Discov*. Data Just Accepted (February 2024). https://doi.org/10.1145/3649506

Ye, H., Liu, T., Zhang, A., Hua, W. et Jia, W. (2023). Cognitive Mirage: A Review of Hallucinations in Large Language Models. *arXiv*:2309.06794v1 [cs.CL]

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, AT, Bi, W., Shi, F. et Shi, S. (2023). Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *arXiv* cs.CL eprint 2309.01219, https://doi.org/10.48550/arXiv.2309.01219

---

i « Un texte non original peut en outre avoir l'air humain pour une raison embarrassante : la répétition et la combinatoire aveugles de patrons ne sont en aucun cas l'apanage des machines. La pensée, la parole et l'écriture humaines sont souvent beaucoup moins authentiques que nous ne voudrions l'admettre. Comme l'a fait remarquer Heidegger (2010), une grande partie de ce que les gens font est fait parce que c'est ainsi que 'l'on' fait les choses. » (Durt et al., 2023 : 11)

ii Il convient de préciser qu'au moment des expériences, les grands modèles de langues entrainés sur des données médicales en français CamemBERT-bio (Touchent et al., 2023) et DrBERT (Labrak et al., 2023) n'étaient pas disponibles. Copara et al. (2020) avait pré-entrainé le modèle CamemBERT sur 31 000 articles scientifiques médicaux en français issues de PubMed. Pourtant, le modèle n'a pas été rendu public.

iii https://github.com/ibuhnila/refomed

iv RefoMed est constitué d'un total de 11653 paires de termes médicaux – reformulations médicales dont 8626 paires de termes médicaux – reformulations médicales sont en français et 3027 paires en roumain. Les données en roumain ont été extraites de GrandMedRo2 (Buhnila, 2023), corpus qui contient des textes de vulgarisation médicales du web.

v Nous avons inclus dans le calcul les prédictions correctes et partiellement correctes.

vi Certaines prédictions ont été difficiles à annoter à cause de leur forme inventée ou artificielle. Les annotateurs n'ont pas pu annoter en fonctions 10 prédictions correctes et 5 incorrectes et en relations 7 prédictions correctes et 7 incorrectes. Ces cas n'ont pas été pris en compte dans les statistiques.

vii « Lorsqu'un mot est acquis au fil de ses usages à l'oral ou à l'écrit, il se charge cumulativement des contextes et co-textes dans lesquels il est rencontré, et notre connaissance de ce mot comprend le fait qu'il est le cooccurrent de certains autres mots dans certains types de contextes. Les séquences de mots construites à partir de ces mots sont à leur tour chargées des contextes et co-textes dans lesquels elles apparaissent. » (Hoey, 2005 : 8)

Annexe 1

Nous avons évalué les générations de prédictions pour la liste de test avec des métriques de l'état de l'art dans le domaine de la similarité sémantiques des textes, BLEU (Post, 2018) et ROUGE (Lin, 2004). Le score BLEU calculé sur des uni-grammes est très important s'il y a un mot ou plusieurs mots en commun entre le *Truth* (la reformulation attendue) et *Prediction* (toutes les prédictions générées pour un terme). Dans notre exemple du **Tableau 9**, la F-mesure calculée pour le score BLEU est de 0,40 pour toutes les prédictions, car il a identifié le terme médical « SMLE » et « SML-E » comme mots similaires, alors que les prédictions sont assez variées.

Tableau 9. Exemple de prédictions générées automatiquement par le modèle T5.

Terme médical:
syndrome myasthénique de Lambert-Eaton
Truth:
(SMLE)
Prediction:
maladie latente
(par exemple apathie ménoadio-nasique)
(syndrome myasthétique de Lambert-Eaton)
(SML-Eaton, syndrome de Hamaslin)
(SGLE)

Dans le **Tableau 10** nous montrons un exemple d'évaluation de la prédiction avec le score ROUGE (utilisant la racinisation), dont la première colonne est ROUGE1 (calculé sur les uni-grammes, nous affichons la F-mesure), la deuxième ROUGE2 (calculé sur les bi-grammes), et ROUGE-L qui calcule la séquence la plus longue de mots communs entre la référence (*Truth)* et la prédiction. Dans notre exemple, la F-mesure de 0.6000 du score BLEU montre que plusieurs mots ont été trouvés en commun avec toutes les prédictions. ROUGE1 indique qu'il y a des mots communs entre la référence et la prédiction, mais qu'ils ne sont pas identiques (des racines communes sont présent *l'urine/urines, urine/urinaire*), ROUGE2 prend en compte de bi-grammes et ROUGE-L la séquence la plus longue des mots communs. Le score ROUGE prend en compte la racinisation des mots, par conséquent, il va considérer les prédictions avec hallucinations (mots inventés) assez similaires et avoir un bon score.

Tableau 10. Exemple de l'évaluation avec le score ROUGE d'une prédiction générée automatiquement.

Terme médical:			
dérivation urinaire			
Truth:	*BLEU*		
(chirurgie pour dériver l'urine de la vessie)	Tensor (0.6000)		
Prediction:	ROUGE1	ROUGE2	ROUGE-L
(définie comme les quantités de urines totales de sang approuvées pour chaque urinaire)	tensor(0.2400)	tensor(0.1600)	tensor(0.1600)

Le BERTscore (calculé avec BERT (Devlin et al., 2018) permet de calculer le cosinus de similarité entre *Truth* et *Prediction* et comparer les contextes d'apparition (**Tableau 11**).

Tableau 11. Exemple d'évaluation avec le score BERTscore.

Terme médical:	
fièvre	
Truth:	*BERTscore*
tue les microbes	
Prediction:	
est la l'infection aigu la plus courante sur la côte nord-est de l'Afrique en 2002	F1=0.971 P= 0.976 R=0.965
/ maladie maladie du foie	F1=0.973 P=0.975 R=0.970
(fièvre, tachycardie, aigu gastro-intestinale	F1=0.967 P=0.973 R=0.961
maladie des fièvres	F1=0.970 P=0.979 R=0.962
(fièvre et fasciologie)	F1=0.965 P= 0.977 R= 0.953