

Journée Néologismes
ATILF, Axe Lexique

16 mai 2025, Nancy



Les néologismes médicaux sont-ils bien reformulés par l'IA générative ?

Ioana Buhnila

Université de Lorraine – CNRS, ATILF UMR 7118

ioana.buhnila@univ-lorraine.fr

Les néologismes

La néologie = la formation de mots nouveaux ou de sens nouveaux

Les néologismes s'intègrent plus ou moins facilement / rapidement dans la langue (parfois pas du tout)

Les néologismes



La néologie = la formation de mots nouveaux ou de sens nouveaux

Les néologismes s'intègrent plus ou moins facilement / rapidement dans la langue (parfois pas du tout)



La néologie = la **génération** (plus aléatoire) de nouvelles formes de **tokens** ou des « **hallucinations** »

Les néologismes sont/ne sont pas intégrés dans le **corpus de pré/post-entraînement** du modèle de langue

Types de néologismes

- Mot existant avec un sens nouveau (néologie de sens)
- Mot ou expression étrangère (emprunt)
- Mot complètement nouveau (néologie de forme)
- Sigles ou acronymes
- Analogie
- Métonymie
- Dérivation

Types de néologismes



- Mot existant avec un sens nouveau (néologie de sens)
- Mot ou expression étrangère (emprunt)
- Mot complètement nouveau (néologie de forme)
- Sigles ou acronymes
- Analogie
- Métonymie
- Dérivation



- Nouveaux tokens dans le corpus de pré ou post entraînement

Les néologismes du COVID : Projet AdaptMed

Objectif du projet AdaptMed

- La création d'un système d'adaptation automatique des textes scientifiques à destination du grand public
- Réaliser une base de données de termes médicaux et leurs reformulations
- Annotation de reformulations et de leur catégorie

Partenariats

- Bibliothèque Nationale de France (DataLab)
- Université de Strasbourg (LiLPa) (Amalia Todirascu)
- Bibliothèque nationale universitaire de Strasbourg (DataLab)
- Soutien du DataLab (BNF)

AdaptMed : Contexte et motivation (I)

Volume important d'informations de vulgarisation scientifique publiées dans les médias, sur les réseaux sociaux

- Variété de reformulations (Fuchs, 2020 ; Eskhol-Taravella et Grabar, 2017) : définition, paraphrase, exemplification
- Difficultés à repérer les mots et les termes complexes

Peu d'outils et de ressources pour la détection des reformulations ou pour l'adaptation dans un domaine cible

- Corpus spécifiques au domaine médical : CLEAR (Cardon et Grabar, 2018) pour la simplification
- Corpus multilingues de paraphrases : PPDB (Gantikevitch et Callison-Burch, 2014)
- Reformulation : Corpus oraux (Eshkol-Taravella et Grabar, 2017), RefoMed (Buhnla, 2023)

AdaptMed : Contexte et motivation (II)

Nouveaux termes médicaux

- « Selon la définition *du Covid long* par l'OMS, "plus de 200 symptômes différents qui affectent la vie quotidienne ont été rapportés". » (France Info, Santé, 24 janvier 2023)
- « *patients atteints d'une forme sévère du Covid-19* », « *plaintes cognitives* », « *biomarqueur positif* » (France Info, Santé, 24 janvier 2023)

Dictionnaires (p.e.Larousse 2023)

"cas contact", "fatigue pandémique", "Covid long", "passe sanitaire et passe vaccinal", "vaccinateur-trice", "commerce essentiel", "Prêt garanti par l'Etat (PGE)".

Diffusion des termes dans les médias

- « le sous-variant BA.5 d'Omicron »
- « biomarqueurs »
- « des anomalies du métabolisme cérébral »

La reformulation et la paraphrase

La reformulation = est définie comme le processus de réécriture qui a le rôle d'expliquer, reprendre ou simplifier une phrase ou un élément de la phrase (Fuchs, 2020 ; Eskhol-Taravella et Grabar, 2017)

La paraphrase = l'équivalence basée sur un noyau sémantique commun (Fuchs, 1982)

Corpus AdaptMed : Collecte de données

1. Sélection des sources (archives du Web Covid –19 de la BNF)
2. Création d'un corpus de reformulations à partir des archives du Web Covid-19 de la BNF.
 - Exemples : *covidose, paucisymptomatique, quatorzaine, super spreader, anosmie, gestes barrière*, etc. (DiCovid-19 (<https://dicovid19.com/>))
3. Définition des requêtes incluant des termes du DiCovid-19 et des marques de reformulation :
 - `text:"masque N95" ~5 AND (collections:"épidémie Covid-19")`
 - `text:"immunogène c'est" ~5 AND (collections:"épidémie Covid-19")`
 - `(text:"coronavirus maladie*" ~5) OR (text: "covid* maladie" ~5) OR (text:"covid* affection?" ~5) AND (collections:"épidémie Covid-19")`
4. Requêtes sur Solr : envoi de requêtes et extraction des phrases

Corpus AdaptMed : Annotation manuelle

- Définition d'un guide d'annotation (Buhnla, 2023)
- Identifier les reformulations
- Identifier les relations lexicales (synonymie, hyponymie, hyperonymie, méronymie)
- Identifier les fonctions sémantico-pragmatiques (définition, explication, paraphrase, exemplification) (Eshkol-Taravella et Grabar, 2017)

Statut	Phrase avec terme et marqueur / corpus BNF2	Terme	Marqueur	Reformulation	Relations lexicales	Fonctions sémantico-pragmatiques
oui	En laboratoire, on arrive à stabiliser certains virus, notamment celui de la dengue.	virus	notamment	dengue	hyponymie	exemplification
oui	Mais il faudra encore attendre car il n'y a pas encore eu d'essai clinique, c'est-à-dire in vivo.	essai clinique	c'est-à-dire	in vivo	synonymie	paraphrase
oui	Mais il faut être prudent, le Covid-19 est un virus nouveau dont le corps n'a pas encore appris à se défendre.	le Covid-19	est une	virus nouveau	hyperonymie	paraphrase

AdaptMed : Génération de reformulations médicales

Contexte expérimental de l'étude (2023)

- **Réseau de neurones APT** (pour la génération de paraphrases)
(Nighojkar and Licato, 2021)
- **Modèle de langue : t5-base (248M)** (multilingue, dont le français)
(Raffel et al., 2020)

AdaptMed: Résultats de génération

Predictions	Nb termes	Pourcentage
Au moins 1 résultat correct	197	34,81 %
Aucun résultat correct	369	65,19 %
Total	566	100 %

Term: COVID longue



- **Truth:** symptômes qui perdurent au-delà de 12 semaines
- **Prediction:** (COVID chronique)

Term: le dépistage



- **Truth:** la réalisation d'examens de détection ou de surveillance
- **Prediction:** une intervention médicale effectuée pour éliminer la maladie

Projet AdaptMed avec t5-base (2023)

=>

Expériences de génération avec des modèles de
langue récents

(2025)

LLMs et SLMs (I)

Les **LLMs** sont des modèles d'apprentissage profond par réseaux de neurones capables de générer du texte automatiquement.

- des résumés (Liu et al. 2024; Song et al. 2024)
- des essais (Tian et al., 2024)
- des textes narratifs (Simon and Muise, 2022)

Petits Modèles de Langue
= *Small Language Models (SLMs)*

Grand Modèles de Langue
= *Large Language Models (LLMs)*

LLMs et SLMs (II)

SLM < 8 milliards de paramètres (8B) > LLM << Llama-3.1 (405B)

Pourquoi utiliser des SLMs ?

- Ils sont adaptés à une application sur des machines moins puissantes et sur des téléphones mobiles.
- Ils ont un temps de réponse plus court et un coût plus faible de calcul.
- Ils permettent une meilleure reproductibilité.

Petits Modèles de Langue (SLMs)

Contexte expérimental de l'étude (2025)

Trois modèles de langue de 3 milliards de paramètres, pré-entraînés sur la langue française, en libre accès.



llama-3.2 (Dubey et al., 2024)



qwen-2.5 (Yang et al., 2024)



phi-3.5 (Abdin et al., 2024)

L'IA générative et les néologismes du Covid

C'est quoi un
cas contact ?



Un "cas contact" est un terme utilisé dans le domaine de la santé publique et des maladies, notamment dans le contexte de l'épidémie de COVID-19.

Un cas contact désigne une personne qui a eu un rapport direct ou indirect avec une personne infectée par le virus SARS-CoV-2, la cause de la maladie du Covid-19. [...]



L'IA générative et les néologismes du Covid

Explique-moi
brièvement c'est quoi
un **cas contact** ?



Un cas contact désigne une personne qui a eu un lien direct ou indirect avec une personne infectée par le virus SARS-CoV-2, comme des membres de la famille, des voisins ou des collègues. Il est surveillé pour détecter les symptômes du Covid-19 et empêcher sa propagation.



L'IA générative et les néologismes du Covid

C'est quoi un super spreader?



Un "super spreader" est une personne qui, en raison de leur exposition à un grand nombre de personnes ou de leur comportement social, transmet rapidement le virus SARS-CoV-2 ou d'autres maladies à une grande quantité de personnes. [...]



Un "super spreader" est une personne qui transmet plus facilement ou plus fréquemment une infection qu'une personne moyenne lors d'une épidémie, notamment dans le contexte des maladies infectieuses. Cette notion a été particulièrement utilisée au cours de la pandémie de coronavirus (Covid-19).[...]



Un "super spreader" est une personne qui, pour des raisons souvent non entièrement comprises, transmet plus fréquemment ou efficacement le virus à d'autres par rapport aux individus typiques dans la même population.

Étude de cas : les néologismes et l'argot

Néologisme argotique : **chokbar**

= choqué (argot internet)

= dérivé de choqué avec un suffixe similaire à -ard (Wiktionary, 2023)

= utilisé par les jeunes à l'oral ou sur les réseaux sociaux

il est chokbar en voyant le pied — (Tweet de @eaudespate, posté le 26 août 2023, à 8h17)

Étude de cas : les néologismes et l'argot

C'est quoi
chokbar ?



Désolé, je n'ai pas trouvé d'informations sur "chokbar" qui correspondent à un terme ou une expression spécifique. Il est possible que c'est un mot ou une expression locale qui n'a pas de traduction directe en français.
[Est-ce un nom d'une marque, un type de produit, un terme populaire ou autre chose ?]



"Chokbar" semble ne pas être un terme couramment utilisé ou reconnu dans les langues que je connais directement, comme le français, l'anglais, etc. Cela pourrait être : *[un nom propre, un terme local ou régional, une erreur typographique]*



Le terme "chokbar" n'a pas de signification largement reconnue en français ou dans un contexte général. Il est possible que cela soit une faute de frappe, un jargon spécifique à un domaine particulier, ou un terme qui a été inventé récemment. Voici quelques possibilités : *[Faute d'orthographe, Termes spécifiques au domaine ou slang, Jargon anglais]*



>>> c'est quoi chokbar?

Désolé, je n'ai pas trouvé d'informations sur "chokbar" qui correspondent à un terme ou une expression spécifique. Il est possible que c'est un mot ou une expression locale qui n'a pas de traduction directe en français.



Pouvez-vous me donner plus de contexte ou des informations supplémentaires sur ce que vous entendez par "chokbar" ? Est-ce un nom d'une marque, un type de produit, un terme populaire ou autre chose ?



>>> c'est un argot

Je comprends mieux maintenant. "Chokbar" est un argot utilisé en France, notamment dans les régions du Nord et de l'Est.



Selon certaines sources, "chokbar" pourrait désigner un type de cocaïne ou une substance illégale similaire à la cocaïne. Cependant, je n'ai pas trouvé de définition officielle ou d'informations précises sur ce terme en argot.

Il est possible que "chokbar" soit un terme local ou régional spécifique qui ne serait pas largement connu ou utilisé ailleurs en France. Si vous avez plus de détails ou des informations sur l'utilisation du terme "chokbar", je serais ravi de les écouter !



>>> c'est un argot utilisé par les jeunes

Je comprends mieux maintenant. Les argots sont souvent spécifiques à certaines régions, milieux ou groupes d'âge, et peuvent varier en fonction du contexte et de la culture.



Il est possible que "chokbar" soit un terme utilisé par les jeunes dans certaines régions, mais je n'ai pas trouvé de définition officielle ou d'informations précises sur son sens ou sa signification.

Si vous avez une idée de ce que pourrait signifier "chokbar" pour les jeunes qui l'utilisent, pourriez-vous me donner des indices ou des informations supplémentaires ? Est-ce un terme qui désigne quelque chose de spécifique, comme un produit ou une activité ?

Je suis là pour aider et échanger, mais je dois également respecter la vie privée et le monde des jeunes. Si vous êtes prêt à partager plus d'informations, je serai ravi de les écouter !

Conclusions sur les résultats des SLMs

- Les néologismes pour les SLMs sont des mots absents ou peu fréquents dans leur corpus d'apprentissage
- Les néologismes argotiques sont difficiles à définir pour un SLM (peu fréquents, utilisation surtout à l'oral et sur les réseaux sociaux)
- Les modèles de langues multilingues sont principalement centrés sur la langue et la culture anglaise
- Les performances des SLMs peuvent être améliorés avec des données externes et des différentes méthodes (*finetuning, RAG, ICL, CoT, agents LLM*)

Références (sélection)

- Abacha B. A., & Demner-Fushman, D. (2019). A question-entailment approach to question answering. *BMC bioinformatics*, 20, 1-23.
- Buhnila I., Sinha A., & Constant M. (2024). [Retrieve, Generate, Evaluate: A Case Study for Medical Paraphrases Generation with Small Language Models](#). In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024) @ACL 2024*, pages 189–203, Bangkok, Thailand. Association for Computational Linguistics.
- 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>
- Buhnila I. (2023). *Une méthode automatique de construction de corpus de reformulation*. Linguistique. Université de Strasbourg, 2023. Français. (NNT : 2023STRAC006). (tel-04226255)
- 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, pp. 149-184.
- Eddine, M. K., Tixier, A. J. P., & Vazirgiannis, M. (2020). Barthez: a skilled pretrained french sequence-to-sequence model. *arXiv preprint arXiv:2010.12321*.
- Fuchs C. (1982). La Paraphrase. *PUF*. Paris, 184 pages.
- Fuchs Catherine. (2020). Paraphrase et reformulation : un chassé-croisé entre deux notions. *Autour de la reformulation*, 36, Droz, pp. 41-55, *Coll. Recherches et Rencontres*, 978-2-600-06051-6.
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2.
- Grabar N., & Eshkol-Taravella I. (2016). Prédiction automatique de fonctions pragmatiques dans les reformulations. In *TALN 2016: Traitement Automatique des Langues Naturelles*, pp. 1-15. Paris, France: hal-01426814.
- Grabar N., & Cardon R. (2018). CLEAR - Simple Corpus for Medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pp. 3-9. Tilburg, the Netherlands: Association for Computational Linguistics. <https://doi.org/10.18653/v1/W18-7002>.
- Labrak, Y., Bazoge, A., Morin, E., Gourraud, P. A., Rouvier, M., & Dufour, R. (2024). BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. In *62th Annual Meeting of the Association for Computational Linguistics (ACL'24)*.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in neural information processing systems*, 33, pp.9459-9474.
- Liu Y., Shi K., He K., Ye L., Fabbri AR., Liu P., Radev D., & Cohan A. (2024). On learning to summarize with large language models as references. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8639–8656.
- Nighojkar, A. Licato, J. (2021). Improving Paraphrase Detection with the Adversarial Paraphrasing Task. *Proceedings of the 59th Annual Meeting of ACL and IJC NLP*, p. 7106–7116, Online. Association for Computational Linguistics.
- Raffel, C., Shazeer N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Wei L., Liu, P. (2020) Exploring the Limits of Transfer Learning with a UnifiedText-to-Text Transformer, *Journal of Machine Learning Research* 21 (2020) 1-67, <https://arxiv.org/pdf/1910.10683.pdf>
- Todirascu, A., Padó, S., Krisch, J., Kisselew, M., & Heid, U. (2012). French and German Corpora for Audience-based Text Type Classification. In *LREC (Vol. 2012)*, pp. 1591-1597).

Merci pour votre attention

Des questions ?

