ANALYSE DES NEOLOGISMES AFFIXAUX PAR LE BIAIS DE SKETCH ENGINE

*Katarína **Chovancová**, Université Matej Bel à Banská Bystrica, katarina.chovancova@umb.sk*

*Lucia **Ráčková**, Université Matej Bel à Banská Bystrica, lucia.rackova@yahoo.com*

Original scientific paper
DOI: 10.31902/fll.51.2025.9
UDC: 811'373.43:004

Résumé : Le présent article s'intéresse à l'utilisation de l'outil de pointe en matière de corpus *Sketch Engine* créé en 2004 à l'Université Masaryk de Brno, République tchèque dans le cadre d'une recherche sur les néologismes. Nous allons éclaircir les points forts et faibles de ce site web en nous servant de notre propre expérience de recherche en néologie lexicale, notamment sur le lexique de la période de la propagation du virus SARS-Cov-2 (2020 – 2022). L'échantillon de la recherche sur la crise sanitaire a été élaboré à partir du quotidien *Libération*. Pour ce faire, nous allons présenter brièvement le logiciel d'origine tchèque et par suite procéder à l'application de ses fonctions dans la recherche en question. Comme le *Sketch Engine* favorise l'analyse quantitative et le recueil des données statistiques, il est à supposer que le principal défi pour l'utilisateur consiste à traiter des données de nature qualitative, particulièrement au niveau sémantique de la langue. Même si le logiciel, grâce à la possibilité de travailler avec des corpus vastes, montre de nouvelles tendances dans la création des mots en français, le manque de lemmatisation, d'étiquetage des affixes et de références temporelles est à constater. Pour les langues à l'écriture plus développée, y compris le français, le travail des chercheurs est rendu encore plus difficile vu l'ortographe contenant des accents qui complique la recherche des entrées. Les résultats scientifiques obtenus peuvent être utiles aux lexicographes pour une recherche et un traitement plus efficaces des nouvelles unités lexicales, ainsi que pour l'amélioration des méthodologies de création de dictionnaires, et ce, également aux linguistes de corpus et autres spécialistes travaillant sur les langues et les lexiques variés.

Mots clés : Libération, Sketch Engine, Covid, néologismes, affixation

1. Introduction

Dans le domaine de la linguistique de corpus, de nombreux logiciels facilitent aujourd'hui le travail des chercheurs, leur permettant de gagner du temps pour se concentrer sur l'analyse des données et l'interprétation des résultats. L'un de ces outils est *Sketch Engine*, un logiciel de pointe, créé en 2004 et largement utilisé en lexicographie. D'origine tchèque, *Sketch Engine* est conçu pour travailler avec des corpus textuels à grande échelle et constitue une base essentielle pour les recherches en linguistique de corpus. « Sketch Engine (SkE) est un logiciel qui récupère des esquisses de mots (word sketches), les regroupe sur la base de relations grammaticales et crée des thésaurus à partir du corpus » (Sketch Engine, 2023).[25] En plus, dans *Sketch Engine*, la version étendue du langage formel CQL (Corpus Query Language) permettant des requêtes complexes est utilisée (Chalupníková et Volková 4).

Notre objectif primordial dans cette recherche est de mettre en relief les qualités et les défauts de l'utilisation du logiciel *Sketch Engine* dans le cadre d'une recherche sur le potentiel créatif du lexique émergeant pendant la période de la propagation du virus SARS-Cov-2 (2020 – 2022).

Les retombées pratiques de cette recherche résident dans la capacité à aider les chercheurs à déterminer si *Sketch Engine* est l'outil approprié en fonction des objectifs et de la nature de leurs recherches. Il est à supposer que le principal défi pour l'utilisateur réside dans le traitement des données qualitatives, notamment au niveau sémantique de la langue. Une question qui demeure est celle du traitement des affixes dans le cadre de cette recherche, ainsi que de leur identification et leur tri à l'aide de cet outil.

Dans cet article, nous nous appuyons donc sur notre propre recherche antérieure réalisée à l'Université Matej Bel de Banská Bystrica, dont les résultats sont présentés notamment dans *La créativité lexicale dans le temps de la pandémie du COVID* (Jesenská, Ráčková et Veselá, Berlin - Bruxelles - Chennai - Lausanne - New York - Oxford : Peter Lang, 1-284) dans le cadre du projet de recherche VEGA n. 1/0748/21 *Le potentiel lexicogénétique du discours médiatique sur la crise* dirigé par Chovancová. Nous nous y inspirons également de travail de Ráčková et Schmitt (47-60). Parmi les travaux effectués par les chercheurs de l'Université Masaryk de Brno et leurs collaborateurs (*inter alia* Kilgarriff,

[25] « Sketch Engine (SkE) je software, který vyhledává slovní profily (*word sketches*), sdružuje je na základě gramatických relací a vytváří z korpusu tezaury » Accédé le [26 octobre 2023].

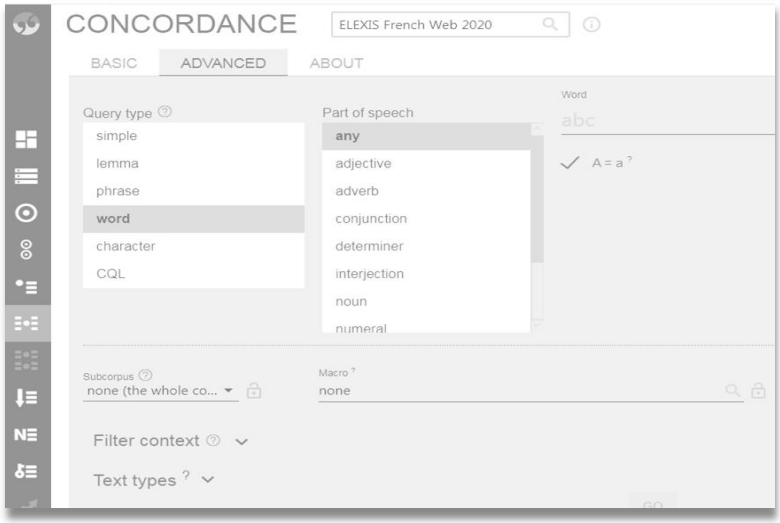
Rychly, Smrz et Tugwell 297-306; Kilgariff et al. 7-36) dont le logiciel est originaire, il y a également une étude qui s'intéresse à l'acquisition de l'anglais à l'aide de la linguistique de corpus (Kilgarriff, Marcowitz, Smith et Thomas 61-80). Les chercheuses espagnoles León-Araúz et Reimerink avec le chercheur québéquois San Martín (893-901) se sont penchés sur le lexique environnemental en travaillant avec un corpus enregistré dans SkE *EcoLexicon English Corpus* (EEC). En outre, le chercheur britannique Pearce (1-29) a décrit son expérience avec *British National Corpus* traité par SkE. Une large utilisation du logiciel pour des langues variées est confirmée par l'étude portant sur l'extraction des collocations grammaticales en chinois (Huang, Kilgarriff et al. 48-55). Une étude intéressante sur le phénomène de la déterminologisation a été publiée par les collègues tchèques, Honová et Holeš (65-77). À la différence de ces études précédentes qui se voient laudatives envers le logiciel de l'analyse textuelle, nous nous permettons non seulement de relever les points forts du logiciel mais aussi de critiquer les points faibles de *SkE*.

Pendant la crise sanitaire mondiale, une épidémie lexicale a éclaté simultanément (Lardellier 87). Cela se reflète dans de nombreuses études analysant le lexique lié au Covid dans une perspective interlinguistique comparative (Jacquet-Pfau et Kacprzak 1-15 ; Dincă 1-15) ou unilingue française (Maldussi 1-20 ; Grimaldi 1-13 ; Labelle et Rondeau 1-16). Pour alléger l'atmosphère pesante, certains chercheurs se sont intéressés aux expressions ludiques créées pendant cette période sans précédent (Guo-Gripay, Berbinski et Veleanu 1-18 ; Tallarico 1-22). Dans le cadre des expressions phraséologiques ou idiomatiques, on retrouve également l'étude de Rollo (1-19). En revanche, Vicari (1-17), toujours dans la même veine, se concentre sur la vulgarisation scientifique des termes médicaux.

En regroupant les sujets relatifs à la néologie lexicale et aux méthodes de linguistique de corpus, cette étude complète d'autres travaux centrés sur le lexique lié au Covid d'une part, et d'autres recherches utilisant *Sketch Engine* comme outil primordial d'autre part (en plus des recherches précédemment mentionnées, celles de San Martín, Trekker et León-Araúz 264-298). De plus, il existe quelques rares études qui associent l'étude du lexique covidien et l'utilisation parallèle de *Sketch* Engine (Rossi 77-94 ; Maurer 1-67 ; Ráčková et Schmitt (47-60)).

2. Brève présentation du logiciel *Sketch Engine*

Un grand atout du logiciel *Sketch Engine* élaboré à l'Université Masaryk de Brno est qu'il : « propose de nombreux corpus prêts à l'emploi, ainsi que des outils permettant aux utilisateurs de créer, de télécharger et d'installer leurs propres corpus » (Kilgariff et al. 7).[26] Ses fonctions principales *Sketch*[27], *Thesaurus*[28], *Concordance*[29] sont aisées à utiliser. La fonction *Concordance* est particulièrement remarquable parce qu'elle sert à voir les utilisations des lexèmes dans des contextes variés. La tâche *GDEX* (*good examples*) permet de générer des exemples-types et ressemble à des dictionnaires en papier qui nous proposent des occurrences des lemmes dans des phrases courtes. De plus, il s'agit en principe d'un logiciel étiqueté, lemmatisé où les utilisateurs peuvent faire leur recherche par rapport à plusieurs catégories, y compris les parties du discours. Ces dernières peuvent être repérées notamment à l'aide de la recherche avancée de la fonction *Concordance* où dans la catégorie du mot, il est possible de choisir parmi leurs catégories traditionnelles, cf. le graphique suivant :



Graphique 1: Concordance et recherche avancée.

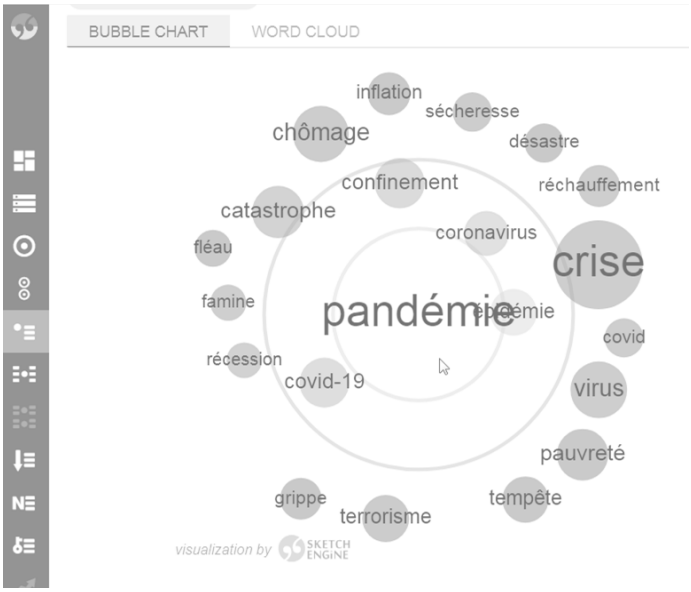
[26] « The Sketch Engine website offers many ready-to-use corpora, and tools for users to build, upload and install their own corpora ».

[27] Résumé d'une page du comportement grammatical et collocationnel d'un mot. Cette fonction a donné le nom au logiciel.

[28] Il s'agit ici d'un outil distributionnel qui montre les collocations.

[29] Un outil de base pour tous ceux qui travaillent avec le logiciel qui permet la recherche des lexèmes dans des contextes variés.

Une qualité incontestable du site web du *Sketch Engine* est la possibilité de créer automatiquement les diagrammes liés à la recherche. Ces visualisations graphiques à portée de tous sont simples à effectuer grâce à une automatisation élevée du logiciel. Grâce à la fonction *Thesaurus*, pour le mot *pandémie*, nous sommes arrivées à un double graphique :



Graphique 2 : Double graphique du lexème « pandémie ».

Il s'agit d'un graphique à bulles illustrant la fréquence des mots dans le corpus *LIBÉ*, que nous avons créé à partir des textes sélectionnés du journal *Libération*. Plus la taille de la bulle est grande, plus la fréquence de l'unité lexicale est élevée dans le corpus. Ainsi, on constate la prédominance des termes « crise » et « chômage » dans le contexte de l'unité lexicale centrale, à savoir « pandémie ».

En revanche, un inconvénient de *Sketch Engine* est qu'il n'est accessible que dans le mode en ligne. De plus : « Le Sketch n'est pas open source, car cela pourrait compromettre sa viabilité en tant qu'entreprise, une version de ce moteur, NoSketchEngine, est open source » (Kilgariff et al. 31).[30] Cela signifie que les créateurs du logiciel proposent une version gratuite d'une durée d'un mois. Bien qu'elle puisse sembler insuffisante, elle peut néanmoins répondre à certains

[30] « While the Sketch Engine is not open source, as this could undermine its viability as a business, a version of it, *NoSketchEngine*, is open source ».

besoins des linguistes, notamment des étudiants ou des doctorants débutants.

3. Recherche en néologie lexicale. Traitement statistique

Avant toute chose, il convient de distinguer les néologismes de forme des néologismes de sens, ainsi que les figures de rhétorique et les néologismes d'emprunt. Souvent, il ne s'agit pas de créer de nouvelles formes de lexèmes, mais d'actualiser le sens des unités lexicales déjà existantes. Dans ce cas, on parle de néologismes de sens ou de néologie sémantique (Cartier, Sablayrolles et al. 1-20). La création des néologismes et leur intégration dans le lexique se réalisent de manière diverse. C'est pourquoi, il s'avère essentiel de prendre en compte le statut de la nouveauté et le statut de leur usage. Dans la présente étude, nous nous intéressons particulièrement aux nouvelles unités émergentes, créées par dérivation affixale.

La dérivation par affixe, également appelée affixation, consiste à ajouter un préfixe ou un suffixe à la base d'un mot. À la préfixation et à la suffixation s'ajoute la parasynthèse, qui consiste en l'ajout simultané d'un préfixe devant la base et d'un suffixe après celle-ci. La dérivation parasynthétique est donc une combinaison de la préfixation et de la suffixation. Nore (2022)[31] en donne un exemple avec le verbe *dératiser*. Selon Corbin (177 dans Cartier, Sablayrolles et al. 19), ce type de dérivation résulte d'une double affixation non simultanée.

Lorsqu'on aborde les processus de formation des mots, notamment l'affixation, il est important de noter qu'une grande partie des unités lexicales en français ne se sont pas formées directement au sein du français, mais ont été empruntées au latin, où elles ont été créées par des procédés dérivationnels. C'est, entre autres, le cas du verbe *revenir* (Petraş 72). De ce fait, la consultation du dictionnaire étymologique est essentielle.

Dans les lignes qui suivent, nous allons montrer comment le logiciel de *SketchEngine* nous a servi dans la recherche en néologie lexicale. En fait, il constituait une base pour atteindre notre objectif, à savoir l'identification et l'analyse du potentiel créatif des affixes, préfixes et suffixes, dans le discours médiatique sur la crise sanitaire.

Le traitement statistique de notre échantillon de recherche, à savoir le corpus LIBÉ contenant dans un seul document 1 922 910 tokens, 1 630 574 mots et 61 846 phrases, s'est déroulé en plusieurs étapes : 1. l'enregistrement des textes sélectionnés dans le logiciel, 2. l'établissement de la liste des 10 mots-clés (cf. Tableau 1) précédé par

[31] [consulté le 26 août 2023].

la création automatique des 20 mots-clés (cf. Graphique 3), 3. la création des listes des unités lexicales affixées grâce aux étoiles de Kleene, 4. la distinction des affixes des unités libres, 5. la classification des néologismes à l'aide du nombre absolu des unités procuré par SkE.

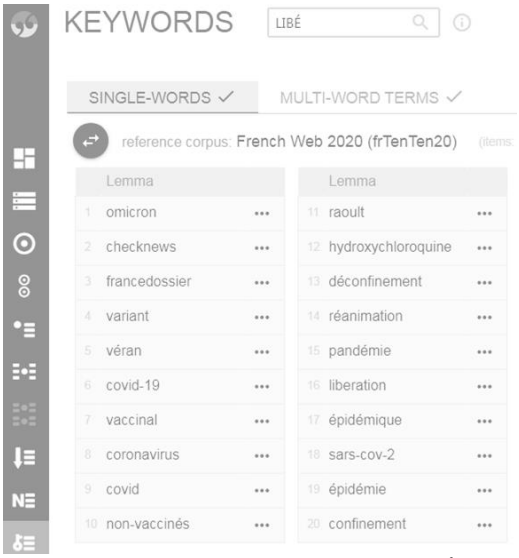
Nous avons travaillé avec une sélection de textes marqués le plus par la crise sanitaire, donc tirés du quotidien pour la période des années 2020 – 2022. De nouveaux lexèmes ou les lexèmes aux sens actualisés dans cette période-là ont été examinés dans leur ensemble, mais également dans leurs trois vagues pandémiques couvrant les étapes suivantes : 1) mars – mai 2020 ; 2) novembre 2020 – janvier 2021 ; 3) novembre 2021 – février 2022.

Pour se lancer dans le vif du sujet, la création de notre corpus de recherche a été la première tâche à effectuer. Pour ce faire, le *SkE* nous a aidées grâce à sa simplicité d'enregistrement des textes des chercheurs, dans ce cas issus du quotidien *Libération* de la période respective. Les textes sélectionnés, réunis dans un seul document *Word* par notre étudiante Viktória Veltyová, scientifique adjointe, ont été archivés par nous-même en utilisant *WebBootCaT*. Le problème de ce type d'enregistrement est le manque de regroupement des unités lexicales selon certains critères, c'est-à-dire en lemmatisation. Comme un codage de différentiation n'était pas appliqué, il nous fallait enregistrer plus de textes, donc le grand corpus *LIBÉ,* et après classer les textes dans des sous-corpus contenant les trois vagues pandémiques. Ce souci était également signalé par notre collègue, Petra Jesenská, travaillant sur l'anglais, forcée à s'abonner à plus d'espace de logiciel.[32]

La création de la liste des mots-clés à l'aide de la fonction *Wordlist* nous a servi en tant que point de départ pour une recherche plus approfondie. À première vue, la présence du phénomène de la vulgarisation scientifique était à noter. Nous avons retenu les termes de médecine : *hydroxychloroquine, vaccinal, non-vaccinés, réanimation* et de virologie : *coronavirus, covid-19, épidémie, épidémique, pandémie, variant, omicron et SARS-Cov-2.* Il fallait cependant se rendre compte que, dans la liste des mots-clés, un tri des néologismes et des unités lexicales déjà établies en lange française se montrait nécessaire. Dans cette optique, nous nous sommes appuyées notamment sur les critères étymologiques et celui de la nouveauté du néologisme potentiel en consultant les dictionnaires électroniques *Trésor de la langue française*

[32] Par ailleurs, la recherche sur l'anglais comme langue d'investigation linguistique est plus simple du point de vue de l'absence des accents. Pour en savoir plus sur le lexique covidien en anglais traité par *Sketch Engine*, voir l'article de Jesenská (18-23).

informatisé et *Dictionnaire de l'Académie française*. Par conséquent, les mots-clés *CheckNews, Francedossier, Libération, Raoult, Véran*, le lexème remontant du XXe siècle *réanimation* et l'abréviation du nom du médicament *hydroxychloroquine* ne faisaient pas partie des unités néologiques.



Graphique 3: 20 mots-clés du corpus *LIBÉ*.

À partir de cette liste des mots-clés du corpus *LIBÉ* (cf. Graphique 3), nous avons ensuite élaboré notre propre sélection des 10 lexèmes-clés (cf. Tableau 1). Celle-ci inclut, entre autres, des lexèmes tels que *crise, épidémie* et *vaccin*, qui sont des unités néologiques lorsque leur utilisation se rapporte à la crise du coronavirus, à l'épidémie de SARS-CoV-2 et aux vaccins spécifiques utilisés contre la COVID-19, tels que *AstraZeneca* ou *Pfizer*. Cette nouvelle liste est également munie du nombre de fréquence pendant les trois vagues pandémiques qui est à observer dans le tableau suivant:

Lexème-clé	1ère vague	2e vague	3e vague
confinement	1683	215	82
coronavirus[33]	1774	231	139
Covid-19	2259	321	1061
crise	1791	204	194
distanciation	141	7	7
épidémie	1408	198	225
pass sanitaire	0	0	190
télétravail	119	11	72
variant	2	175	657
vaccin	176	518	629

Tableau 1 : 10 Lexèmes-clés pendant les trois vagues pandémiques.

Grâce à la fonction de *SkE* affichant le nombre absolu des occurrences pour chaque unité lexicale, nous avons pu établir des listes de la fréquence des *candidats néologismes*[34]. Ce classement par importance a construit un terrain propice à l'identification des tendances actuelles du français. Néanmoins, en ce qui concerne la reconnaissance des affixes, préfixes et suffixes, leur distinction des unités libres reste incontournable. Par exemple, une grande partie des unités lexicales en français ne se sont pas formées au sein du français, mais ont été directement reprises du latin où elles se sont formées par les procédés de dérivation. Dans le corpus *LIBÉ*, nous avons relevé les unités lexicales préfixales à la fréquence élevée. Nous les présentons dans le Tableau 2 qui confirme la productivité du préfixe *anti-*. À part *anticorps*, ce préfixe a donné naissance à de nouvelles unités lexicales telles que *Anticovid*, *anticoronavirus*, *anti-coronabonds*, *anti-confinement*. Le préfixe *sur-* fait preuve d'une productivité réalisée importante en exprimant l'idée de superiorité comme dans les unités lexicales « surmortalité » ou « survie » :

[33] Le nombre total des unités lexicales *coronavirus* et *Covid-19* n'égale pas le nombre du corpus *LIBÉ* car le logiciel du *Sketch Engine* n'arrive pas à distinguer différentes formes orthographiques. Dans les 8 unités lexicales restantes, le nombre total des unités des trois vagues égale au nombre total du corpus *LIBÉ.*

[34] terme de Cartier, Sablayrolles et al. 9

Préfixe	Unité lexicale	Nombre d'occurrences absolu
anti-	anticorps	178
in-	incertitude	109
sur-	surmortalité	64
	survie	70

Tableau 2 : Unités lexicales à emploi fréquent et très fréquent formées par la préfixation.

Quant à la suffixation, il est à constater que les suffixes font preuve d'une grande variabilité. La productivité réalisée se manifeste notamment dans les suffixes 1. *–isme, -iste*, *-ment* et 2. *–age, -(at)ique*. La première triade des suffixes montre un phénomène intéressant, il s'agit des substantifs nouvellement créés à partir des verbes, donc des déverbatifs, à savoir *alarmisme, antivaxinisme, multilatéralisme / complotiste, conspirationniste, covidiste / confinement, déconfinement, isolement*[35]. Il semble qu'un autre suffixe productif en formation des substantifs, se montre le suffixe *–age* : *décalage, séquençage*. Et enfin, un suffixe prolifique pour la création des adjectifs est *-(at)ique*: *épidémique, symptomatique.*

Suffixe	Unité lexicale	Nombre d'occurrences absolu
-age	décalage	53
	séquençage	39
-(at)ique	épidémique	265
	symptomatique	55
-isme	alarmisme	4
	antivaxinisme	1
	multilatéralisme	12
-iste	complotiste	12
	conspirationniste	5
	covidiste	1

[35] Le lexème *isolement*, tout comme les lexèmes crise, pandémie et vaccin, est un néologisme de sens et non un néologisme de forme. Cela signifie que l'unité lexicale *isolement* a simplement élargi son sens pour être utilisée dans le contexte de la crise du Covid.

-ment	confinement	1981
	déconfinement	480
	isolement	208

Tableau 3 : Les néologismes créés par la suffixation.

C'était grâce aux quantificateurs, astérisques, appelés aussi « étoiles de Kleene » (Chalupníková et Volková 7), que nous avons pu rechercher des unités lexicales qui commencent par un préfixe, par exemple *dé* ou qui finissent par un suffixe, par exemple *age*. Malgré cette fonction-là et son apparente simplicité, nous sommes arrivées à une liste exhaustive des données où tout d'abord, il fallait distinguer les affixes d'éléments initiaux des mots où le sens compositionnel (cf. Jalenques 39 ; Petraş 62-75), important pour l'identification du préfixe, n'était pas perçu. Et dans une deuxième étape, il était indispensable d'effectuer un tri manuel minutieux des nouveaux lexèmes des mots courants, voire des mots fréquemment employés dans le discours médiatique sur la crise sanitaire, en français, la langue de cette investigation linguistique.

Un autre problème émergeant pour nous, chercheuses intéressées par le français, était la question des accents. À la différence de *Google Search*, le site web de *Sketch Engine* ne trouve que des unités lexicales tapées avec des accents bien placés dans la fonction *Concordance.* Cela peut être un inconvénient également pour les chercheurs désirant distinguer différentes formes orthographiques. Quant à nous, nous avons voulu savoir combien d'occurrences pour les variations Covid-19[36] covid-19 et COVID-19 se manifestaient dans le corpus. La seule solution à ce problème-là se révélait être un tri manuel minutieux.

4. Discussion des résultats

Après une brève présentation du logiciel *Sketch Engine* et de notre expérience de chercheuses, il reste pourtant quelques points à discuter. L'espace du logiciel est en théorie facile à utiliser. Néanmoins, pour les personnes avec des aptitudes techniques insuffisantes, son utilisation peut certainement poser des problèmes. Il est aussi préférable que les utilisateurs maîtrisent l'anglais, surtout pour la communication avec l'équipe gérant le logiciel en cas de problèmes, mais aussi pour son utilisation étant donné qu'il est la première langue de l'interface.

[36] Forme prédominante finalement.

L'une des fonctions du logiciel qui à notre avis laisse à désirer est la fonction *Keywords* permettant de générer les mots-clés. Dans notre corpus constitué des textes du quotidien *Libération*, et par suite dénommé *LIBÉ*, la liste des 20 mots-clés créée automatiquement par le logiciel était étrange. Comme nous avons pu l'observer au-dessus, elle comportait des dénominations étonnantes comme *Checknews* [37] ou *Francedossier*[38] à quoi s'ajoutaient des noms de famille comme *Raoult, Véran* et le nom du quotidien *Libération*. Nous pouvons affirmer que, bien que ces lexèmes soient indéniablement fréquemment utilisés dans le texte, ils ne constituent en aucun cas les mots-clés du corpus en question. Nous avons, par conséquent, dû fabriquer notre propre liste des mots-clés, plus adaptée à notre recherche et à ses objectifs.

Force est de constater que le logiciel *Sketch Engine* présente un caractère ambivalent. D'une part, il est extrêmement utile pour les linguistes de corpus, permettant de repérer les unités lexicales et d'étudier leurs caractéristiques formelles. D'autre part, il est difficile d'identifier les qualités sémantiques en raison de l'absence d'étiquetage des séquences de mots, telles que les affixes. En conséquence, il est impossible de mener une analyse sémantique sans un tri manuel considérable des unités lexicales et sans consulter leurs occurrences spécifiques dans les dictionnaires et dans leur contexte. Dans le domaine de la recherche en sémantique lexicale, il est donc essentiel de toujours remettre en question les résultats obtenus avec *SkE.* Cependant, ce défaut est partiellement compensé par la possibilité d'effectuer une analyse contextuelle approfondie, notamment grâce à la fonction *Concordance.*

5. Conclusion

Le logiciel de *Sketch Engine* se révèle être un outil très avantageux pour donner aux chercheurs une première image générale de leur corpus et pour travailler avec ce corpus-là. Il s'avère cependant insuffisant pour une recherche plus approfondie. Son défaut le plus important reste, à notre avis, le manque de lemmatisation. L'étiquetage des affixes est problématique et apparemment, il y a un long chemin à faire, même dans les outils automatisés.

Par conséquent, même si notre corpus contenait des dates précises, pendant son enregistrement dans le logiciel, il n'était pas possible de les introduire dans le corpus et de savoir après pendant

[37] Un nouveau type de moteur de recherche géré par des journalistes lancé par le quotidien *Libération* pour mieux comprendre l'actualité.
[38] La rubrique de *Libération* consacrée à l'actualité en France.

quelle vague pandémique ces néologismes-là ont apparu. Pour cela, il fallait faire des sous-corpus pour la première, deuxième et troisième vague pandémique. Par contre, certains corpus de *SkE* contiennent une annotation temporelle, à savoir *French Web 2020 (frTenTen20)* qui montre la date d'accès sur la page web ainsi que le site web où les énoncés ont été repérés, par exemple *EUR-Lex 2/2016 parallel – French* contient une référence temporelle minimale, année du document.

Ce qui nous a toutefois agréablement surprises, c'est la richesse des corpus disponibles, non seulement en français, mais également dans des langues moins diffusées, notamment le slovaque. Parmi ceux-ci figurent *Araneum Slovacum*, *DGT – Translation Memory Parallel – Slovak* et *ELEXIS Slovak Web 2021*. C'est précisément sur cette langue que portera notre future recherche, centrée sur la suffixation dans une perspective comparative franco-slovaque.

Nous espérons que les développeurs de Sketch Engine prendront conscience des limites du logiciel et l'actualiseront en fonction des besoins de ses utilisateurs. En effet, ces derniers ne se contentent pas seulement d'observer les lexèmes dans des contextes variés, mais souhaitent également avoir accès à des dates précises, des entrées lemmatisées, ainsi qu'à des données permettant de saisir les entrées sans accent, notamment pour les langues ayant des systèmes orthographiques plus complexes.

Quant au lexique du Covid, il représente encore un vaste champ à explorer, particulièrement en ce qui concerne la disparition et la préservation des néologismes liés à la crise sanitaire récente.

6. Remerciements

Cette contribution est publiée dans le cadre du projet de recherche n° 09I03-03-V04-00417 financé par l'Union européenne NextGenerationEU par le biais du Plan de relance et de résilience de la République slovaque.

Funded by the
European Union
NextGenerationEU

Références bibliographiques :

Académie française. *Dictionnaire de l'Académie française.* 9e éd., https://www.dictionnaire-academie.fr. Accédé le 10 septembre 2023.

Cartier, Emmanuel, Sablayrolles, Jean-François et al. « Détection Automatique, Description Linguistique et Suivi des Néologismes en Corpus: Point d'Étape sur les Tendances du Français Contemporain. » *Congrès Mondial*

de Linguistique Française (CMLF), SHS Web of Conferences, vol. 46, article 08002 (2018) : 1-20. https://doi.org/10.1051/shsconf/20184608002.

Chalupníková, Daniela, et Nikol Volková. « Dotazy v CQL (pro Sketch Engine), Stručný přehled CQL s důrazem na vyhledávání v českých korpusech. » PLIN021: *Sémantická Analýza v Praxi,* 2013, https://nlp.fi.muni.cz/trac/research/rawattachment/wiki/cs/SketchEngine/Dotazy_v_cql.pdf. Accédé le 8 janvier 2024.

Corbin, Danielle. *Morphologie Dérivationnelle et Structuration du Lexique*. 2 vols, Max Niemeyer Verlag, 1987.

Trésor de la langue française informatisé. *ATILF*, https://atilf.atilf.fr. Accédé le 10 septembre 2023.

Dincă, Daniela. «Lexique Roumain de la Pandémie dans la Communication Institutionnelle » *Repères DoRiF* 25 (2022) : 1-15.

Grimaldi, Claudio. «Les traits de la distance et de l'isolement dans le lexique autour de la pandémie » *Repères DoRiF*, 25 (2022) : 1-13.

Guo-Gripay, Weiwei, Berbinski, Sonia et Corina Veleanu. « La création lexicale de la pandémie, entre peur et humour » *Repères DoRiF*, 25 (2022) : 1-18.

Honová, Zuzana et Jan Holeš « Term in non-specialised context. Case of determinologisation of psychiatric terminology ». *Folia linguistica et litteraria. Journal of Language and Literary Studies* (2023) : 65-77. DOI: 10.31902/fll.45.2023.4.

Huang, Chu-Ren, Adam Kilgarriff, Yiching Wu, Chih-Ming Chiu, Simon Smith, Pavel Rychlý, Ming-Hong Bai et Chen, Keh-Jiann. « Chinese Sketch Engine and the extraction of grammatical collocations ». *Proceedings of the fourth SIGHAN workshop on Chinese language processing* (2005) : 48-55.

Jalenques, Pierre. « Quand la diachronie renvoie à la synchronie : étude des emplois idiomatiques du préfixe re en français (renier, remarquer, regarder etc.) ». *Recherches linguistiques de Vincennes*, 30 (2001) : 39-62.

Jesenská, Petra. « Neological Covid lexis from the viewpoint of word formative process in English *». Lingua Viva : odborný časopis pro teorii a praxi vyučování cizím jazykům a češtině jako cizímu jazyku*, 19/37 (2023) : 18-23, https://www.pf.jcu.cz/images/PF/veda-vyzkum/linguaviva/download/LV37_2023.pdf. Accédé le 2 septembre 2024.

Jesenská, Petra, Lucia Ráčková et Dagmar Veselá. *La créativité lexicale dans le temps de la pandémie du COVID.* Berlin – Bruxelles – Chennai – Lausanne – New York – Oxford : Peter Lang, 2025.

Kacprzak, Alicja, et Christine Jacquet-Pfau. « De quelques mots-témoins d'une pandémie: Les représentations du Covid-19 en français et en polonais » *Repères-Dorif* 25 (2022) : 1-15.

Kilgarriff, Adam, Pavel Rychly, Pavel Smrz et David Tugwell. « The sketch engine ». *Practical Lexicography: a reader* (2008) : 297-306.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovař, Jan Michelfeit, Pavel Rychlý, et Vít Suchomel. «The Sketch Engine: ten years on ». *Lexicography ASIALEX 1* (2014) : 7-36. DOI: 10.1007/s40607-014-0009-9.

Kilgarriff, Adam, Frederik Marcowitz, Simon Smith et James Thomas. « Corpora and language learning with the Sketch Engine and SKELL ». *Revue française de linguistique appliquée 20* (1) (2015) : 61-80.

Labelle, Mélanie, et Karine Rondeau. « Le travail des terminologues en temps de crise: l'expérience du Bureau de la traduction du gouvernement du Canada » *Repères DoRiF* 25 (2022) : 1-16.

Lardellier, Pascal. *Petite anthropologie d'une crise sanitaire.* Paris : MkF Éditions, Coll.Les essais médiatiques, 2022.

Libération 2020 – 2022. *Sketch Engine*, https://app.sketchengine.eu/#dashboard?corpname=user%2Flucia.rackova%2Flibe_2. Accédé le 26 octobre 2023.

León-Araúz, Pilar, Antonio San Martín et Arianne Reimerink. « The EcoLexicon English corpus as an open corpus in Sketch Engine » *arXiv preprint arXiv:1807.05797*. *Lexicography in global contexts* (2018) : 893-901.

Maldussi, Danio. « De nouvelles dénominations pour un concept ancien: le rôle de l'adjectif qualificatif, de l'adjectif relationnel et du substantif épithète dans les processus d'innovation néologique en temps de pandémie » *Repères DoRiF*, 25 (2022) : 1-20.

Maurer, Liina. « Dispersion de la crise qui continue : Covid-19 dans les discours journalistiques. Mémoire de master, 2023.

Nore, Françoise. *Les Néologismes*. 2022, https://www.francoisenore.com/articles/les-neologismes. Accédé le 26 août 2023.

Pearce, Michael. « Investigating the collocational behaviour of MAN and WOMAN in the BNC using Sketch Engine ». *Corpora*, *3*(1), (2008) : 1-29. DOI: 10.3366/E174950320800004X.

Petraş, Cristina. « Archaisme, lexicalisation et variation sur le terrain acadien : autour des verbes en RE/re-/r- et associés ». *Revue de sémantique et pragmatique. Cadrage sur la variation, le changement lexical et le changement grammatical en francais actuel* (2017) : 41-42. DOI : 10.4000/rsp.451.

Ráčková, Lucia et François Schmitt. « Étude sémantique des verbes préfixés en français dans le discours médiatique sur la crise sanitaire ». *Studia Romanistica*, 23/1 (2023) : 47-60.

Rollo, Alessandra. « Métaphores et "covidiomes" aux temps de la Covid-19 *». Repères DoRiF*, 25 (2022) : 1-22.

Rossi, Micaela. « Vulgariser les concepts scientifiques dans la presse: Les définitions par métaphore » *Roczniki Humanistyczne* 70.8 (2022) : 77-94.

San Martín, Antonio, Catherine Trekker, et Pilar León-Araúz. « Repérage automatisé de l'hyponymie dans des corpus spécialisés en français à l'aide de Sketch Engine » *Terminology* 28.2 (2022) : 264-298.

Sketch Engine, www.sketchengine.eu. Accédé le 26 octobre 2023.

Tallarico, Giovanni. « Néologismes expressifs et ludiques dans le vocabulaire de la pandémie ». *Repères DoRiF*, 25 (2022) : 1-22.

Vicari, Stefano. « Quand les médecins deviennent influenceurs: la vulgarisation des termes de la Covid-19 dans Facebook, Instagram et Twitter». *Repères DoRiF, DoRiF*, 25 (2022) : 1-17.

ANALYSIS OF AFFIXAL NEOLOGISMS USING SKETCH ENGINE

This paper focuses on the use of the state-of-the-art corpus tool, *Sketch Engine* (*SkE*), which was created in 2004 at Masaryk University in Brno, Czech Republic, as part of research into lexical semantics. We will discuss the strengths and weaknesses of this tool based on our own experiences as researchers in lexical neology, particularly in studying the lexicon from the period of the SARS-CoV-2 virus spread (2020–2022) with affixation as the focus of research interest. The daily newspaper *Libération* was the primary and sole source for constructing our research sample related to the health crisis, referred to as *LIBÉ.*

Regarding the structure of the article, we will briefly introduce the Czech-origin software and then proceed by applying its functions to our research. As *Sketch Engine* favours quantitative analysis and the collection of statistical data, it can be assumed that the main challenges for users involve processing qualitative data, particularly at the semantic level of language. The scientific results obtained may be useful not only for lexicographers but also for corpus linguists and others working on various languages and lexicons.

Thus, *Sketch Engine*, as a text analysis software, has been in use for just over 20 years. Its primary role is to facilitate the work of researchers, helping them save time in order to focus on analysing data and interpreting results. It is therefore widely used in lexicography and serves as a basis for working with large-scale text corpora. The software retrieves word sketches, groups them based on grammatical relationships, and creates thesauri from the corpus. Additionally, *Sketch Engine* employs an extended version of the formal language CQL (*Corpus Query Language*), allowing for complex queries (Chalupníková and Volková 4).

Our overarching goal in this research is to highlight the qualities and shortcomings of *Sketch Engine* in researching the creative potential of the lexicon emerging during the SARS-CoV-2 virus spread (2020–2022). A practical outcome of this study is to assist researchers in deciding whether or not to use *Sketch Engine* based on the objectives and nature of their research. We assume that the main challenges for users will be processing qualitative data, particularly at the semantic level of language. It remains to be seen how to deal with the affixes in this research and how to identify and sort them using this tool.

In this paper, we draw on our previous research conducted at Matej Bel University in Banská Bystrica, the results of which are presented in, among others, *La créativité lexicale dans le temps de la pandémie du COVID* (Jesenská, Ráčková and Veselá, Berlin - Bruxelles - Chennai - Lausanne - New York - Oxford: Peter Lang, 2025) published as part of the VEGA No. 1/0748/21 research project

The Lexicogenetic Potential of Media Discourse on the Crisis directed by Chovancová. Among the work carried out by researchers at Masaryk University in Brno and their collaborators (Kilgarriff, Rychly, Smrz and Tugwell 297–306; Kilgarriff et al. 7–36), from which the software originates, there is also a study on the acquisition of English using corpus linguistics (Kilgarriff, Marcowitz, Smith and Thomas 61–80). Spanish researchers León-Araúz and Reimerink, along with Quebec researcher San Martín (893–901), focused on the environmental lexicon using a corpus recorded in the *Sketch Engine EcoLexicon English Corpus* (*EEC*). British researcher Pearce (1–29) described his experience with the *British National Corpus* processed by *Sketch Engine.* The wide usage of the software across various languages is further confirmed by studies on extracting grammatical collocations in Chinese (Huang, Kilgarriff et al. 48–55). An interesting study on the phenomenon of determinologisation was published by Czech colleagues Honová and Holeš (65–77).

During the global health crisis, a lexical epidemic broke out simultaneously (Lardellier 87). This is reflected in numerous studies analysing the COVID-related lexicon from a comparative cross-linguistic (Jacquet-Pfau and Kacprzak 1–15; Dincă 1–15) or unilingual French (Maldussi 1–20; Grimaldi 1–13; Labelle et Rondeau 1–16) perspective. To lighten the heavy atmosphere, some researchers focused on the playful expressions created during this unprecedented period (Guo-Gripay, Berbinski and Veleanu 1–18; Tallarico 1–22). Rollo's study (1–19) also deals with phraseological or idiomatic expressions. On the other hand, Vicari (1–17), again in the same vein, concentrates on the scientific popularisation of medical terms.

By grouping together topics relating to lexical neology and corpus linguistic methods, this study complements other works focusing on the COVID-related lexicon on the one hand, and other research using *Sketch Engine* as a primary tool on the other (in addition to the previously mentioned research by San Martín, Trekker and León-Araúz 264–298). In addition, there are a few studies that combine the study of the COVID-related lexicon with the parallel use of *Sketch Engine* (Rossi 77–94; Maurer 1–67; Ráčková and Schmitt 2023).

A key advantage of the *Sketch Engine* website is its ability to automatically create search-related diagrams. These graphical visualisations are easily accessible to all users, thanks to the software's high level of automation. A significant strength of the software is that "[t]he Sketch Engine website offers many ready-to-use corpora, and tools for users to build, upload, and install their own corpora" (Kilgarriff et al. 7). Its main functions – *Sketch* (a one-page summary of a word's grammatical and collocational behaviour), *Thesaurus* (a distributional tool that shows collocations) and *Concordance* (a basic tool for anyone working with the software, which enables lexemes to be searched in a variety of contexts) – are easy to use. The *Concordance* function is particularly noteworthy as it shows how lexemes are used in various contexts. The *GDEX* (*Good Dictionary Examples*) task can generate typical examples, similar to how paper dictionaries display lemma occurrences in short sentences. Furthermore, the program is labelled and lemmatised, allowing users to search across several categories, including parts of speech. These can be located using the advanced

search in the *Concordance* function, where users can select their traditional category in the word category. The software's ability to automatically create diagrams linked to searches, particularly through the *Thesaurus* function, further enhances its usability.

Unlike previous studies, which are generally laudatory (e.g. Pearce 1–29, León-Araúz, Reimerink and San Martín 893–901) regarding the text analysis software, we take the liberty of criticising the weak points of *Sketch Engine* and highlighting the difficulties we encountered in our work as researchers in lexicology, specifically in lexical semantics.

Even though the software is labelled and lemmatised, users can search across several categories, including parts of speech. These can be located using the advanced search in the *Concordance* function. However, when searching for affixes, prefixes and suffixes, the lemmatisation leaves something to be desired. It is impossible to formally distinguish between simple parts of words, their initial parts or endings.

The creation of the keyword list with the help of the *Wordlist* function served as a starting point for further research. At first glance, the presence of the phenomenon of scientific popularisation was noticeable. We selected medical terms: *hydroxychloroquine, vaccinal, non-vaccinés, réanimation* and virology terms: *coronavirus, covid-19, épidémie, épidémique, pandémie, variant, omicron* and *SARS-Cov-2.* However, it had to be realised that, in the list of keywords, a sorting out of neologisms and lexical units already established in French language was necessary. The keywords *CheckNews, Francedossier, Libération, Raoult, Véran*, the twentieth-century lexeme *réanimation* and the abbreviation of the drug name *hydroxychloroquine* were not among the neologistic units. From this list, we then had to create our own list of key lexemes, with their frequency during the three pandemic waves, consisting of the following lexemes: *confinement, coronavirus, Covid-19, crise, distanciation, épidémie, pass sanitaire, télétravail, variant* and *vaccin* whose distribution changed with each of the three pandemic waves.

Sketch Engine provided a basis for achieving our objective – identifying and analysing the creative potential of affixes, prefixes and suffixes in the media discourse on the health crisis. The statistical processing of our research sample took place in several stages: 1) registration of our texts in the software; 2) creation of a list of keywords; 3) creation of lists of affixed lexical units using Kleene stars; 4) distinction of affixes from free units; and 5) classification of neologisms using the absolute number of units provided by *Sketch Engine*.

The software facilitated our work, thanks to its simplicity in recording the researchers' own texts, in our case, the daily newspaper *Libération* from the respective period. The selective texts, collected into a single Word document by our student assistant scientist Viktória Veltyová, were archived by us using *WebBootCaT*. However, the lack of grouping of lexical units according to certain criteria, such as lemmatisation, posed a problem. Since differentiation coding was not applied, we had to record more texts, resulting in a large *LIBÉ* corpus, which we then classified into sub-corpora corresponding to the three pandemic waves.

Nevertheless, when it came to recognising affixes, prefixes and suffixes, distinguishing them from free units was unavoidable. It was thanks to quantifiers – asterisks, also known as Kleene stars (Chalupníková and Volková 7) – that we could search for lexical units starting with a prefix (e.g. *dé*) or ending with a suffix (e.g. *age*). Despite this function's apparent simplicity, we ended up with an exhaustive list of data where we first had to distinguish initial element affixes from words where the compositional meaning (cf. Jalenques 39; Petraş 62–75), important for identifying the prefix, was not perceived. Despite the difficulties encountered, we arrived at several interesting conclusions. For example, the productive suffix in noun formation is shown to be the suffix *-age* (e.g. *décalage, séquençage*), and the productive suffix for creating adjectives is *-(at)ique* (e.g. *épidémique, symptomatique*). It must be noted that while *Sketch Engine* is very useful for corpus linguists to find lexical units and study their formal qualities, it is less effective for semantic analysis, given the lack of labelling of word sequences such as affixes. As a result, semantic analysis requires a significant amount of manual sorting of lexical units.

Additionally, using *Sketch Engine* for an extended period requires a subscription. "While *Sketch Engine* is not open-source, as this could undermine its viability as a business, a version of it, *NoSketchEngine*, is open-source" (Kilgarriff et al., 2014, 31). This means that the software's creators offer a free version for one month, which, although potentially insufficient, may suffice for certain needs, particularly for beginners, students or PhD students. A free subscription is also available through the software's university of origin, Masaryk University in Brno, or with an e-mail belonging to that domain.

The *Sketch Engine* software proves to be a very advantageous tool for giving researchers an initial general picture of their corpus and for working with it. However, it is insufficient for more in-depth research. Even if the software, thanks to the possibility of working with vast corpora, shows new trends in the creation of words in French, the lack of lemmatisation, affix labelling and temporal references is to be noted. For languages with more developed writing, including French, the work of the researchers is made even more difficult by the fact that the spelling contains accents, which complicates the search for entries.

On the other hand, it should be noted that parallel corpora in multiple languages are now available in *Sketch Engine*, giving researchers the option to choose them for comparative research if desired.

What's more, even though our corpora contained precise dates, during its recording in the software it was not possible to introduce them into the main corpus *LIBÉ* and find out afterward during which pandemic wave these texts, and consequently these neologisms, appeared. To do this, it was necessary to create sub-corpora for the first, second and third pandemic waves. However, certain *Sketch Engine* corpora contain a temporal annotation, namely *French Web 2020* (*frTenTen20*), which shows the date of access to the webpage as well as the website where the statements were found. For example, *EUR-Lex 2/2016*

parallel - *French* contains a minimal temporal reference, the year of the document.

In addition, what pleasantly surprised us was the vast number of corpora, not only those in French, but also those for less widely spoken languages, including Slovak, which will be the focus of our future research into suffixation from a comparative French–Slovak perspective. We hope that the developers of *Sketch Engine* will realise the limitations of the software and update it according to the needs of its users. Indeed, users are not content simply to observe lexemes in a variety of contexts, but also want access to precise dates, lemmatised entries and data enabling entries to be entered without accents, particularly for languages with more complex orthographical systems.

As for the COVID lexicon, it still represents a vast field to be explored, particularly as regards the disappearance and preservation of neologisms linked to the recent health crisis.

Keywords: *Libération*, *Sketch Engine*, COVID, neologisms, affixation