STUDIA UBB PHILOLOGIA, LXVII, 1, 2022, p. 235 - 254
(RECOMMENDED CITATION)
DOI:10.24193/subbphilo.2022.1.14

TERMINOLOGIES ÉMERGENTES ET EXPLORATION DE CORPUS SPECIALISÉ. LE LEXIQUE DE LA COVID-19 EN ROUMAIN

CRISTINA VARGA[1]

Article history: Received 11 September 2021; Revised 11 November 2021; Accepted 30 November 2021; Available online 31 March 2022; Available print 31 March 2022

ABSTRACT. ***Emerging Terminologies and Exploration of Specialized Corpus. The Lexicon of Covid-19 in Romanian.*** This study discusses the issue of emerging terminologies analysed through quantitative and qualitative analysis of a specialised corpus. Thus, beginning from studies on terminological neologisms such as Rondeau (1984), Kageura (2002) and Humbley (2009) who state that emergent terminologies are formed as a result of a combination of terminologies already fixed in the language and terminological neologisms. Humbley (2009) also hypothesizes that most of the terminological units of an emerging domain are neologisms. Using this theoretical framework and working with a corpus of more than 400,000 words, we aim to observe the terminology of the emerging field of Covid-19. As objectives of this research we aim to answer the following questions: a) which are the most used monolexical terms of Covid-19 in Romanian, b) which are the most used plurilexical terms of Covid-19 in Romanian, c) is the emerging terminology of this field mostly composed of neological terms? The study is meant for researchers, teachers and students interested in corpus-based linguistic research.

Keywords*: emerging terminologies, corpus linguistics, specialised corpora, neology, glossary*

[1] **Cristina VARGA** est Maître de conférences à la Faculté des Humanités de l'Université Catholique de l'Ouest à Angers et chargé de cours du Département de langues modernes de l'Université « Babeş-Bolyai » de Cluj-Napoca (Roumanie) où elle enseigne Informatique pour traducteur, Outils de traduction assistée par ordinateur, Corpus pour traducteur, Localisation, Traduction audiovisuelle (sous-titrage) et Terminologie. Elle est depuis 2011 professeur invité à Barcelona School of Management de l'Université Pompeu Fabra, où elle enseigne le sous-titrage dans le Master de traduction littéraire et audiovisuelle. Cristina Varga a obtenu son doctorat à l'Université « Babeş-Bolyai » de Cluj-Napoca et à l'Université Pompeu Fabra de Barcelone, sa thèse de doctorat : La transmission des connaissances dans le cyberespace. Analyse du discours des forums Web professionnels en tant que sous-genre de l'Internet. Elle a une riche expérience didactique à l'étranger (France, Belgique et Espagne). Ses domaines de travail et de recherche comprennent : l'analyse du discours, la linguistique de corpus, la création et la gestion des corpus multilingues, la traduction automatique, la terminologie, la traduction audiovisuelle et la localisation. Courriel électronique : cristina.varga@ubbcluj.ro.

REZUMAT. ***Terminologii emergente şi explorarea unui corpus specializat. Vocabularul Covid-19 în limba română.*** În prezentul studiu, se discută problema terminologiilor emergente analizate prin intermediul analizei cantitative şi calitative a unui corpus specializat. Astfel, pornind de la studii despre neologismele terminologice precum Rondeau (1984), Kageura (2002) şi Humbley (2009), care afirmă că terminologiile emergente se formează în urma unei combinaţii dintre terminologiile fixate deja în limbă şi neologismele terminologice, vom analiza terminologia emergentă din domeniul Covid-19 în limba română. De asemenea, Humbley (2009) avansează ipoteza conform căreia cea mai mare parte dintre unităţile terminologice ale unui domeniu emergent sunt neologisme. În acest context, pornind de la un corpus de peste 400.000 de cuvinte din domeniul Covid-19 ne propunem să observăm terminologia acestui domeniu emergent. Ca obiective ale acestei cercetări ne propunem să răspundem la următoarele întrebări: a) care sunt cei mai utilizaţi termeni monolexicali din domeniul Covid-19 în limba română, b) care sunt cei mai utilizaţi termeni plurilexicali din domeniul Covid-19 în limba română, c) este terminologia emergentă a domeniului Covid-19 în limba română formată în majoritate din termeni neologici? Studiul se adresează cercetătorilor, profesorilor şi studenţilor interesaţi de cercetarea lingvistică bazată pe corpus.

Cuvinte-cheie*: terminologii emergente*, *lingvistică de corpus*, *corpus specializat*, *neologie*, *glosar*

1. Introduction

Les domaines émergents et leurs terminologies constituent un sujet d'intérêt pour la recherche terminologique depuis plusieurs décennies. Ils attirent facilement l'attention des locuteurs et des spécialistes en raison des nouvelles unités lexicales spécialisées, des néologismes et des emprunts linguistiques qui les caractérisent, et de la fréquence à laquelle ils sont utilisés dans la communication. Les premières tentatives de distinction et de définition des néologismes terminologiques sont dues aux études de Rondeau (1984, 121-124), développées au fil du temps par d'autres études comme Kageura (2002, 43-162) et Humbley (2009, 3-4). Ce dernier affirme que les terminologies émergentes sont basées sur des terminologies déjà fixées dans la langue. Cette idée se retrouve à la fois chez Kageura (2002) et Humbley (2009), les auteurs affirmant que toute construction de nouvelles connaissances s'appuie sur des connaissances existantes et que tout nouveau domaine s'appuie sur un domaine déjà connu. Les terminologies émergentes se développent donc elles aussi de manière similaire, en se basant sur des termes fixés dans la langue et utilisés dans la communication spécialisée dans différents domaines.

Une autre hypothèse avancée et analysée dans Humbley (2009) fait référence au fait que les domaines spécialisés émergents s'appuient sur une terminologie néologique, qui leur est essentielle. Sur la base de cette étude, nous analyserons dans cet article si la terminologie émergente du domaine Covid-19 en roumain est majoritairement constituée de néologismes terminologiques ou, au contraire, de termes déjà connus et fixés en roumain.

En étudiant la terminologie émergente du domaine Covid-19, on constate que de nombreux chercheurs mettent l'accent sur ses aspects néologiques (A. Roig-Marín 2020; Samylicheva et Gazda 2020; Asif *et alii* 2021) sans mentionner les nombreuses unités terminologiques déjà existantes utilisées dans ce domaine émergent. Dans le présent article, poursuivant une série de recherches terminologiques que nous avons menées au cours des deux dernières années, nous nous proposons d'utiliser la méthode de la linguistique de corpus pour mettre en évidence les unités terminologiques les plus fréquemment utilisées en roumain dans le domaine de la Covid-19, en soulignant, le cas échéant, leur caractère néologique.

Ainsi, à partir d'un *Corpus de textes écrits en roumain portant sur le domaine de la Covid-19* développé par l'auteur de cette étude et basé sur les principes de création et d'implémentation de corpus (L'Homme 2004; Sinclair, 2005; Wynne, 2005; McEnery, Hardie, 2012), nous utiliserons différents instruments électroniques d'exploration et d'interrogation de corpus pour observer et analyser la terminologie émergente de la Covid-19 en roumain et son utilisation dans les documents officiels et les médias en Roumanie.

Les objectifs de cette recherche sont orientés vers l'observation des unités terminologiques monolexicales et polylexicales identifiées dans le corpus ainsi que des mots-clés les plus utilisés dans la communication sur Covid-19 dans les documents officiels et dans les médias roumains. Notre recherche vise à répondre aux questions suivantes : a) Quels termes monolexicaux dans le domaine de la Covid-19 sont les plus fréquemment utilisés dans la communication officielle et les médias en Roumanie ? b) Quels termes polylexicaux sont les plus fréquemment employés dans la communication officielle et les médias en Roumanie ? c) Quels sont les mots-clés les plus fréquemment utilisés dans ce domaine ? d) Est-ce vrai que la terminologie émergente de la Covid-19 est largement néologique ? Ces réponses nous permettront d'établir le profil spécifique de l'emploi de la terminologie de la Covid-19 dans les documents officiels et dans les médias en roumain.

Cette étude s’adresse autant aux spécialistes en linguistique de corpus et aux terminologues comme aux traducteurs, aux interprètes, aux professeurs et aux étudiants intéressés par ce domaine.

2. Méthodologie de la recherche

Afin de mener à bien les objectifs proposés dans le cadre de cette recherche, les différentes étapes du travail ont été rigoureusement planifiées. Ainsi, la première étape, essentielle pour l'ensemble de la démarche, consiste à créer le corpus et à le développer en plusieurs étapes successives. S'agissant d'une étape essentielle de notre recherche, les principes de base du corpus ont été pris en compte lors de sa construction. Il s'agit de : *représentativité, équilibre, thématique, taille, niveau de spécialisation* et *homogénéité* (L'Homme 2004 et Sinclair 2005). Nous avons également dû tenir compte de la spécificité du domaine émergent de la Covid-19 qui est en constante évolution. Les changements qui se sont produits avec la croissance et le développement du domaine doivent également être reflétés dans notre corpus. C'est ainsi qu'a été créé le *Corpus de textes écrits en roumain dans le domaine de la Covid-19*, un corpus spécialisé, développé en trois étapes, chacune d'entre elles illustrant le développement du domaine à un certain stade apportant de nouvelles dimensions de la communication dans l'espace public dans ce domaine.

Des méthodes de travail différentes ont été utilisées lors des étapes de la constitution du corpus. Ainsi, le fait que, dans un premier stade, la collection de textes était relativement petite et que l'inventaire terminologique n'était pas très étendu a permis d'organiser le matériel et d'extraire les termes manuellement, sans utiliser d'outils informatiques. La deuxième étape du développement du corpus, déjà plus complexe et comportant plus de textes, a nécessité l'utilisation d'outils informatiques dans la recherche. Ainsi, au cours de cette étape, la plateforme *SketchEngine (https://www.sketchengine.eu/)* a été utilisée pour enrichir la collection de textes du corpus, pour un contrôle plus efficace des données linguistiques et pour l'identification, la délimitation et l'extraction de la terminologie. Les fonctions de la plateforme virtuelle ont également été utilisées pour effectuer les requêtes nécessaires et obtenir des listes d'unités terminologiques monolexicales, polylexicales et de mots-clés. Le logiciel *AntConc (https://www.laurenceanthony.net/software/antconc/)* a également été utilisé pour des requêtes complémentaires, notamment pour obtenir des données sur la distribution des termes dans le corpus et pour effectuer des concordances.

La troisième étape du développement du corpus a été similaire à la précédente, les méthodes de travail étaient similaires et le corpus a atteint une taille de 417 465 mots. Dans cette dernière étape, les textes officiels et médiatiques relatifs à la campagne de vaccination et les textes mentionnant la quatrième vague de la pandémie ont été intégrés. Il en résulte un corpus de textes écrits en roumain provenant de sources officielles et médiatiques, représentatif de toutes les étapes du développement de la pandémie de Covid-19 en Roumanie.

Après l'élaboration de cette source d'information linguistique dans le domaine de la Covid-19, le corpus a été interrogé afin d'obtenir les informations linguistiques nécessaires à la poursuite de notre étude. Les requêtes ont été effectuées successivement, en commençant par la liste de mots du corpus, d'où ont été extraits les termes monolexicaux, suivie de la création d'une liste de *clusters* et de *collocations*, en utilisant la fonction *N-grams* de la plateforme *SketchEngine*. A partir de la liste générée automatiquement, les 500 premiers résultats obtenus ont été consultés, ordonnés selon leur fréquence dans le corpus, et seuls les termes polylexicaux rencontrés ont été extraits. Enfin, la fonction *Keywords* de la plateforme *SketchEngine* a été utilisée pour générer des phrases et des mots-clés représentatifs du corpus interrogé.

Au total, 300 unités terminologiques identifiées ont été analysées à partir des listes générées automatiquement, dont l'auteur a éliminé les erreurs inhérentes à l'analyse textuelle informatique, les éléments lexicaux fragmentés et les mots appartenant au vocabulaire courant. Au final, on a obtenu trois listes de termes qui forment un inventaire d'unités lexicales spécialisées représentatives du domaine de la Covid-19. Le classement automatique des termes et des syntagmes en fonction de leur fréquence d'utilisation dans le corpus est très utile, car il nous permet de savoir quels sont les termes les plus fréquemment utilisés, s'il s'agit de termes néologiques ou non, et aussi, à partir de ces listes, il est possible d'établir un profil de l'inventaire des termes de la Covid-19 utilisés dans la communication officielle et les médias en Roumanie.

Le classement automatique des termes et des syntagmes en fonction de leur fréquence d'utilisation dans le corpus est très utile, car il nous permet de savoir quels sont les termes les plus fréquemment utilisés, qu'il s'agisse de termes néologiques ou non, et aussi, à partir de ces listes, il est possible d'établir un profil de l'inventaire des termes Covid-19 utilisés dans la communication officielle et les médias en Roumanie.

L'analyse quantitative et qualitative des données linguistiques nous permettra de faire des affirmations empiriques sur l'utilisation des termes Covid-19 dans la langue roumaine et d'observer si, comme l'affirment les chercheurs cités ci-dessus, ce domaine émergent utilise également une terminologie principalement néologique ou, au contraire, s'appuie sur des savoirs existants.

Tous les résultats de l'analyse et toutes les données obtenues à partir de l'interrogation du corpus sont mentionnés dans les conclusions de cette étude.

3. La description du corpus

La création du *Corpus de textes écrits en roumain du domaine de la Covid-19* a été réalisée en plusieurs étapes, chaque étape de développement et d'expansion du corpus étant influencée par l'évolution du domaine Covid-19.

Chaque développement du corpus a également entraîné des changements dans ses paramètres. Ainsi, dans une première phase de développement, il comprend 97 172 mots et se compose de 110 textes authentiques écrits en roumain et publiés sur Internet entre mars 2020 et juillet 2020. Tous les textes qui composent ce premier corpus de base sont issus d'une autorité publique, médicale ou administrative et ont un caractère officiel. La typologie des textes composant le corpus à ce stade, le nombre de textes choisis et leur taille obéissent au principe d'équilibre et de représentativité, essentiel pour un corpus. En termes de structure interne, le corpus est composé de : 20 textes - protocoles officiels activés dans le contexte de Covid-19 (source: Ministère de la Santé), 28 textes - informations pour les citoyens roumains voyageant à l'étranger (source: Ministère des affaires étrangères), 37 textes - informations d'intérêt général sur la pandémie de la Covid-19 (source: Institut national de santé publique), 8 textes - ordonnances militaires sur les mesures de prévention de l'infection par la Covid-19 (source: Police roumaine), 5 textes - recommandations à la population dans le contexte de la pandémie (source: Police roumaine), 1 texte - questions fréquemment posées aux citoyens (source: Police roumaine), 4 textes - informations sur l'infection à Covid-19 (source: Société roumaine de microbiologie), 7 textes - informations d'intérêt général sur l'infection à Covid-19 (source: Medlife, Regina Maria).

Les textes présentent une densité terminologique moyenne et un niveau de spécialisation modéré. En tant que situation de communication, il s'agit d'une communication émanant d'une autorité publique, administrative ou médicale qui transmet des informations générales au grand public afin de l'informer de la crise sanitaire pour qu'il puisse comprendre et s'adapter à la situation et respecter les mesures recommandées.

Cette première phase de développement du corpus correspond à la première vague de la pandémie de Covid-19 en Europe et dure à partir de mars 2020 jusqu'à fin juillet 2020. Par la suite, au fur et à mesure que la situation pandémique évolue, la terminologie du domaine évolue également et le corpus doit donc être complété pour inclure des textes correspondant à cette situation.

La deuxième étape de développement du corpus illustre la communication dans le contexte de la Covid-19 entre août 2020 et janvier 2021. Pendant cette période, la structure du corpus est étendue et enrichie. Le *Corpus monolingue de textes écrits en roumain dans le domaine de la Covid-19* atteint 324 045 mots, 399 402 tokens, 14 919 phrases et 290 textes. Tous ces textes ont été ajoutés au corpus original en utilisant la plateforme *SketchEngine*.

Contrairement au premier ensemble de textes qui a constitué le corpus initial, dans cette deuxième étape, en raison de la généralisation du débat sur les questions de la Covid-19 dans les médias, les textes sélectionnés pour

figurer dans le corpus n'étaient pas seulement des textes officiels issus des autorités locales, mais aussi des textes des médias roumains. La priorité a été donnée aux documents de la presse nationale et locale et aux textes écrits provenant de sites des télévisions. Les textes provenant de réseaux sociaux ou de médias à sensation ont été systématiquement rejetés, car ils ont été jugés comme des sources d'information peu fiables et donc non pertinents pour cette étude. Comme dans la première étape, seuls les textes authentiques écrits en roumain ont été sélectionnés.

La troisième étape de la constitution du corpus est similaire à l'étape précédente, puisque nous utilisons également la plateforme en ligne SketchEngine. Cette étape couvre la période de développement de la terminologie Covid-19 comprise entre février 2021 et juillet 2021 et enrichit considérablement le corpus. Il rejoint une taille de 509 883 tokens, 417 465 mots, 20 008 phrases et 383 textes. Un grand nombre de ces textes proviennent des médias roumains, parmi les URL les plus récurrents figurent : *adevarul.ro, mediafax.ro, ziare.com, realitatea.net, wall-street.ro, digi24.ro, dw.com, stirileprotv.ro* et *cotidianul.ro*. Le domaine virtuel le plus important d'où proviennent les textes qui composent le corpus est .ro, ce qui souligne le fait que ce corpus est correctement structuré car il s'agit de pages web roumaines créées par des parlants natifs du roumain et sur lesquelles sont publiés des textes authentiques en roumain. Les autres noms de domaine mentionnés comme sources des textes sélectionnés à ce stade sont : *com, net, eu, tv* et *org*. Ces noms de domaine indiquent que de nombreux textes sélectionnés font partie de sites web de médias roumains.

Cette dernière étape de l'enrichissement du corpus inclut les textes et la terminologie liés à la campagne de vaccination contre la Covid-19 et la quatrième vague. Même si pendant cette période, un développement terminologique du domaine a pu être observé à travers l'apparition de nouveaux termes, nous considérons que le sujet n'est plus autant dans l'attention du grand public.

Il est également important de noter que, bien qu'aucun texte comportant du contenu sur le sujet de l'anti-vaccination, ou provenant du domaine des théories de la conspiration, n'ait été recherché ou sélectionné, l'analyse du corpus a mis en évidence l'existence de ces concepts dans le contexte du débat sur la pandémie de Covid-19. C'est pourquoi nous avons décidé d'analyser également ces unités terminologiques spécialisées, même si elles ne font pas l'objet de notre étude. Nous pensons qu'ils ont leur importance puisqu'ils sont mentionnés dans le même contexte que les termes les plus importants du domaine Covid-19.

4. Analyse et exploration du corpus

L'analyse et l'exploration d'un corpus spécialisé fournit non seulement le matériel nécessaire pour organiser et documenter un inventaire des termes

d'un domaine dans un glossaire, mais aussi le moyen de mettre en évidence la portée et la fréquence d'utilisation de certains termes dans une langue particulière.

L'exploration du *Corpus de textes écrits en roumain dans le domaine de la Covid-19* permet d'observer comment la terminologie de ce domaine est utilisée en roumain. Les requêtes qui peuvent être effectuées à l'aide de logiciels d'exploration de corpus tels que *SketchEngine* ou *AntConc* peuvent fournir des informations sur la fréquence d'utilisation *des termes monolexicaux, des termes polylexicaux* et *des mots clés* en roumain. Ces données empiriques permettent d'établir un profil spécifique illustrant les aspects et les sujets d'intérêt les plus pertinents dans le domaine de la Covid-19 pour les locuteurs roumains. Dans ce qui suit, nous analyserons ces données, mais non sans avoir distingué plusieurs sous-domaines au sein du domaine Covid-19. Il est déjà connu que l'inventaire des termes de ce domaine est très riche, et le fait de le séparer en sous-domaines nous permettrait de faire des observations linguistiques plus précises et plus pertinentes. *Les termes monolexicaux* et *polylexicaux* les plus employés selon les sous-domaines établis seront également analysés.

Dans le domaine de la recherche terminologique, une méthode très efficace pour documenter et organiser l'inventaire linguistique d'un domaine spécialisé est la création de schémas conceptuels. Ceux-ci permettent au chercheur d'organiser de manière logique tous les concepts appartenant au domaine étudié et d'exclure ceux qui ne forment pas de liens logiques avec les autres composants du système conceptuel.

Dans les recherches terminologiques précédentes[2], l'auteur a développé le schéma conceptuel du domaine Covid-19 pour les termes identifiés, délimités et extraits du corpus en roumain. Ce schéma conceptuel est actuellement accessible à l'URL : *https://www.mindmeister.com/1812434005*. Selon ce schéma conceptuel, le domaine Covid-19 peut être divisé en 8 sous-domaines Covid-19 suivants : *concepts généraux, épidémiologie, diagnostic, pathogénie, traitement, prévention, manifestations cliniques* et *protection*. A ces sous-domaines courants dans la documentation de la terminologie d'une maladie, nous avons ajouté une catégorie spécifique de termes provenant d'un sous-domaine moins courant : *le scepticisme vaccinal et les théories de la conspiration*, car des termes appartenant à cette catégorie de discours ont été observés lors des requêtes, même si aucun matériel de ce type n'a été sélectionné pour notre corpus. Nous désirons déterminer, à l'aide de données quantitatives et qualitatives, dans quelle mesure ces termes sont généralement présents dans la communication Covid-19.

2 Il s'agit d'articles en cours de publication qui ont été présentés lors du 20e Colloque international du département de linguistique : Langue roumaine - modernité et continuité dans la recherche linguistique (Bucarest, 20-21 novembre 2020)/ The 20th international conference of the department of linguistics: romanian language – modernity and continuity in linguistics research (Bucharest, november 20–21, 2020) et à l'occasion de la XVI Giornata scientifica realiter terminologia e interculturalità. problematiche e prospettive, 1 e 2 ottobre 2020, organisée par l'Université de Bologne et l'Université catholique du Sacré-Cœur, Milan.

Par conséquent, suite à l'interrogation du corpus, un nombre de 300 termes a été initialement retenu, qui, après une évaluation attentive, la liste a été réduite à 259 termes. La liste est formée par les *termes monolexicaux* les plus fréquemment utilisés dans le *Corpus de textes écrits en roumain du domaine de la Covid-19*. Selon le sous-domaine auquel ils appartiennent, les termes ayant la plus grande fréquence d'utilisation appartiennent au sous-domaine de la *Prévention*, qui contient 67 termes, suivi du sous-domaine de *l'Épidémiologie* avec 56 termes. Le sous-domaine le moins représenté est celui des *Manifestations cliniques*, avec uniquement 9 termes fréquemment utilisés. La situation complète selon le nombre de termes les plus fréquents des sous-domaines de la Covid-19 dans les documents officiels et les médias en Roumanie peut être consultée dans le tableau suivant :

Sous-domaine	**Nombre de termes**
Prévention	67
Épidémiologie	56
Général	38
Protection	28
Pathogénie	27
Traitement	18
Diagnostic	16
Manifestations cliniques	9

Cette répartition des termes Covid-19 les plus fréquents selon les sous-domaines suggère que le débat public, tant au niveau officiel que médiatique, a été principalement orienté vers la discussion des causes et de la nature de la pandémie de Covid-19, mais surtout vers les moyens de la prévenir et de la combattre. On constate également que les aspects plus spécialisés de la pandémie, représentés par des termes se référant au traitement, au diagnostic et aux manifestations cliniques de la maladie, ne sont pas aussi fréquents dans le débat public.

4.1. Termes monolexicaux

Les requêtes à l'échelle du corpus visant à identifier les *termes monolexicaux* dans le domaine Covid-19 ont donné lieu à la création d'une liste de tous les mots constituant le corpus, affichés en fonction de leur fréquence, à l'aide de la fonction *Wordlist* de la plateforme *SketchEngine* et du programme *AntConc*. En raison du fait que la réponse générée par les deux logiciels était une liste de 31 886 mots, dont une partie seulement était des *termes monolexicaux* appartenant au domaine étudié, il a été décidé de limiter l'analyse aux 300 premiers termes rencontrés et à leurs formes, liste qui a ensuite été réduite à 259 termes ayant la plus grande fréquence dans le corpus.

De même, les termes extraits du corpus ont été conservés tels qu'ils étaient dans le texte source, sans aucune intervention. Ainsi, pour un même terme, différentes formes flexionnelles peuvent être observées dans les résultats des requêtes du corpus. Même dans le cas des textes, très nombreux, qui n'ont pas été écrits avec des diacritiques roumains, la forme des mots a été préservée sans être altérée. Les formes d'un même terme ont bien sûr été regroupées tout au long de l'analyse pour illustrer non seulement la fréquence d'un terme dans le texte mais aussi la variété des formes avec lesquelles il est utilisé dans la communication en roumain.

Ainsi, le terme le plus fréquemment rencontré dans le corpus est *covid-19* qui a une fréquence de 2 202 occurrences, tandis que le terme ayant la plus faible fréquence parmi ceux sélectionnés est *désinfectants*, avec 42 occurrences dans le corpus, selon *SketchEngine*[3]. Ce résultat de l'interrogation du corpus confirme que le corpus suit les principes qui sous-tendent la création d'un corpus (L'Homme 2004; Sinclair 2005 et McEnnery, Hardie, 2012) puisque l'hyperonyme du domaine du corpus est également le terme le plus fréquent. Il s'agit de vérifier que la manière dont le corpus a été créé et sa structure sont optimales pour l'analyse des faits de langue en question. De plus, le fait que la liste des termes sélectionnés contienne des termes appartenant à tous les sous-domaines mentionnés confirme que le corpus est équilibré et représentatif.

Une brève analyse des dix premiers *termes monolexicaux* ayant la plus grande fréquence d'utilisation dans la langue roumaine dans le contexte de la Covid-19, nous permet d'affirmer quels sont les concepts les plus intéressants pour les autorités locales et les médias en Roumanie en matière de communication dans ce domaine. Ainsi, les 10 premiers termes et variantes terminologiques dont la fréquence est la plus élevée sont : *covid-19, vaccinare, vaccin, doze, vaccinarea, coronavirus, vaccinul, vaccinate, pfizer, cazul.* Parmi ceux-ci, on note le nom de la maladie et de l'agent pathogène qui a déclenché la pandémie comme principaux sujets de discussion, suivis par des termes appartenant au même champ lexical : *vaccinare, vaccin, vaccinarea, vaccinul, vaccinate*. Pas moins de 5 items lexicaux faisant référence soit à l'action de *vaccination*, soit à la modalité de prévention, *le vaccin*. Deux autres termes à très haute fréquence se retrouvent dans le même champ lexical, à savoir le terme *doze* et *pfizer*, c'est-à-dire « doze de vaccin » et un type de vaccin. Par ailleurs, le terme *cazul* illustre la préoccupation du débat public concernant les cas d'infection à Covid-19 dans la population.

3 Dans cette étude, nous avons donné la priorité à la plateforme virtuelle *SketchEngine*, sur laquelle le corpus est enregistré. Le logiciel *AntConc* a également été utilisé de manière constante tout au long de l'étude pour le contrastage des données et pour effectuer des requêtes complémentaires, notamment pour observer la distribution d'un terme dans le corpus analysé, une fonction qui n'existe pas sur la plateforme *SketchEngine*.

À l'autre extrémité du spectre se trouvent les termes les moins utilisés de la liste des 259 termes sélectionnés pour l'analyse. Cela ne signifie pas qu'il s'agit de termes de basse fréquence. Ils continuent de figurer parmi les termes à haute fréquence du corpus, mais pour notre analyse, ils présentent une fréquence inférieure à celle des termes déjà examinés. Ce sont : *injecție, internat, mănuși, procedura, restricții, sănătate, vaccinarii, vaccinezi, asimptomatice* et *dezinfectanți*. Et dans le cas de ces termes ayant une fréquence plus faible, nous pouvons observer trois termes liés au champ lexical du vaccin : *injecție, vaccinarii* et *vaccinezi*. La plupart des autres termes relèvent du domaine de la prévention et de la protection médicale : *internat, mănuși, procedura, restricții* et *dezinfectați*. Cette liste comprend également des termes généraux tels que *sănătate* et *asimptomatice*.

Sur la base des données de fréquence d'utilisation, nous estimons qu'il est possible de dire que les discussions officielles et médiatiques sur Covid-19 en roumain concernent principalement *vaccinarea* et autres moyens de prévention et de protection de la population.

Pour une analyse plus précise, on peut examiner les unités terminologiques monolexicales issues de l'interrogation du corpus en fonction des sous-domaines Covid-19 mentionnés ci-dessus. A cette fin, les termes les plus fréquents et leurs formes tels que mentionnés dans le corpus ont été organisés dans le tableau suivant. Notez que tant les sous-domaines que les termes et leurs formes sont ordonnés en fonction de leur fréquence observée dans le corpus. Comme il s'agit de formes morphologiques différentes, les occupations n'ont pas été agrégées. Ainsi, on peut voir par exemple que dans le sous-domaine *Prevenție*, la forme du pluriel *vaccinări* n'est pas aussi fréquente dans le corpus que *vaccinare*. Ou bien, on peut également observer qu'en roumain, dans la communication officielle et médiatique sur Covid-19, une plus grande importance est accordée à l'action de *vaccinare* qu'à la forme *vaccinului*.

Sous-domaines	Termes et variantes terminologiques
Prévention	vaccinare vaccinarea vaccinate vaccinărilor vaccinării vaccinări vaccinarii
	vaccin vaccinul vaccinului vaccinuri vaccinurile vaccinurilor

Sous-domaines	Termes et variantes terminologiques
	doze doza
	adverse
	centrul centre centrele centrelor

La préoccupation pour la vaccination devient immédiatement évidente si l'on examine les cinq termes les plus fréquents dans le domaine *Prevenție.* Pratiquement tous les termes appartiennent au champ lexical de la vaccination : *doze*, *adverse* et *centru*. Ils se réfèrent à *doze de vaccin*, *reacții adverse* et *centre de vaccinare*.

En termes d'épidémiologie, nous constatons que la discussion se concentre sur les victimes de l'infection par Covid-19, représentée dans le texte par diverses références telles que : *cazuri, pacienți* ou *contact*. Les termes *confirmat* et *risc* peuvent faire référence à *caz confirmat* ou à *risc epidemiologic*.

Epidémiologie	cazul caz cazurile cazurilor
	pacienții
	confirmate confirmat
	risc
	contact

En termes de *concepts généraux*, il en existe une grande diversité, mais les plus visibles dans les discussions dans les médias sont le nom de la maladie et l'agent pathogène qui a déclenché la pandémie : *Covid-19* et *coronavirus*, suivis des noms commerciaux des différents types de vaccins actuellement utilisés pour prévenir la maladie: *pfizer, moderna* et *astrazeneca*.

Concepts généraux	covid-19
	coronavirus coronavirusului
	pfizer
	moderna
	astrazeneca

Parmi les moyens de protection les plus populaires dans le contexte de la pandémie, dans la communication officielle et dans les médias en langue roumaine, les plus fréquemment mentionnés font référence à *masca de protecție, măsură de izolare* et aux solutions désinfectantes à base de *clor* et *alcool*.

Protecție	protecție[4] protecție protecția protectie protecția
	mască masca măşti măştii
	izolare izolarea
	clor
	alcool

Dans le contexte de la *pathogénie*, on constate que même si *varianta delta* du virus est un terme récent, il présente la fréquence la plus élevée du corpus dans ce sous-domaine. Les autres termes du même champ lexical sont : *virus* et *tulpină*, illustrant une préoccupation significative pour les mutations des coronavirus. Des termes tels que *infectat* et *risc* complètent le panorama des débats dans ce sous-domaine.

Pathogénie	delta
	virus virusuri
	tulpina
	infectate
	riscul riscului

On peut affirmer qu'avec le sous-domaine *tratament* nous nous rapprochons de la spécialisation médicale. C'est également un sous-domaine de la Covid-19 qui a fait l'objet de nombreuses controverses, certains professionnels de la médecine prétendant avoir mis au point des traitements viables pour la Covid-19, ce qu'aucun d'entre eux n'a démontré à ce jour. Dans le présent corpus, on constate que les termes les plus fréquemment utilisés ne sont pas des termes spécialisés, mais des termes généraux: *tratament, medicamente, ser* et *terapie*. Néanmoins, les nombreuses occurrences du terme *antibiotice* renvoient aux diverses controverses et discussions sur leur efficacité dans le cas de Covid-19. Ainsi, dans plusieurs pays, et pas seulement en Roumanie, l'autorité publique a dû intervenir et signaler leur inefficacité afin de prévenir une automédication erronée parmi le grand public.

[4] Dans cette étude, toutes les variantes graphiques du corpus ont été prises en considération : celles sans diacritiques et celles avec diacritiques utilisant les différentes normes pour les claviers roumains SR 13411:1999, ISO/IEC 8859 16:2001 et les diacritiques avec cédille.

Traitement	tratamentul tratament tratamentului
	medicamente medicamentului medicament medicamentelo medicamentele
	serul
	terapie terapia
	antibiotice

Dans le domaine du diagnostic médical, une préoccupation importante est, comme le montre la fréquence d'utilisation du terme, la possible présence des *reacţii* dans le contexte de la vaccination. Tous les autres termes enregistrés comme ayant une fréquence très élevée dans ce sous-domaine appartiennent au champ lexical du test contre la Covid-19. Ce sont : *probe, teste, testare* et *anticorpi*.

Diagnostic	reacţii
	probe probei
	testele testul
	testare
	anticorpi anticorpii

En ce qui concerne les symptômes de la maladie, les plus mentionnés dans le corpus roumain sont : *febră, hipersensibilitate, tuse* et *temperatură*. Étonnamment, Covid-19 symptômes spécifiques : *anosmia* et *ageuzia* ont extrêmement peu d'occurrences dans l'ensemble du corpus.

Manifestations cliniques	simptome
	febră
	hipersensibilitate
	tuşiţi
	temperaturi temperatura

D'autres termes qui attirent l'attention par la richesse des variantes présentes dans le corpus et qui ont une très grande fréquence d'utilisation seraient : *a vaccina, imunizare, infecţie, variantă* et *boală*.

Une catégorie spéciale, avec une présence modeste dans la communication officielle et dans les médias en Roumanie, par opposition aux réseaux sociaux

où elle a une visibilité beaucoup plus grande, est formée par les termes faisant référence aux *scepticisme vaccinal* et *aux théories de la conspiration*.

Par conséquent, un inventaire complet des termes de cette catégorie existant dans le corpus analysé a été réalisé. Bien que les documents aient été sélectionnés de manière à ce qu'il n'y ait pas de textes portant sur les thèmes du scepticisme vaccinal ou des théories de la conspiration, on peut constater que la discussion sur ce sujet a atteint les médias et qu'elle présente même une grande variété terminologique sur le plan de la forme et de l'orthographe.

Le tableau ci-dessous montre uniquement les termes employés dans le corpus. Dans la plupart des cas il s'agit d'une occurrence unique, car ce sont des termes qui apparaissent occasionnellement dans la communication médiatique en Roumanie.

Scepticisme vaccinal et théories de la conspiration	anti-vaccin
	anti-vaccinare
	anti-vaxxerilor anti-vaxxer-ilor anti-vaccinişti
	teorii ale conspiraţiei teoriile conspiraţioniste teoriile conspiraţiei
	conspiraţionismului conspiraţionism
	cospiraţionist antivaxxer conspiraţionistul
	mesajului conspiraţionist
	conspiraţie
	scepticismul antivaccin
	covidioţii conspiraţionişti şi antivaccinişti

4.2. Termes polylexicaux

Les termes polylexicaux sont beaucoup plus difficiles à identifier, à délimiter et à extraire d'un corpus à l'aide de requêtes. Les algorithmes sur lesquels reposent les fonctions d'interrogation sont en grande partie basés sur l'analyse des occurrences des *clusters*, des *collocations* et de leur fréquence d'utilisation dans le corpus. La liste générée automatiquement n'enregistre pas que des termes, au contraire, l'examen d'une telle liste peut faire apparaître beaucoup de « bruit ». Ainsi, l'interrogation des 500 premiers résultats générés par la fonction *N-grams* à partir du *Corpus de textes écrits en roumain dans le domaine de la Covid-19* a un résultat modeste car seulement 36 termes polylexicaux ont pu être identifiés.

Les termes identifiés étaient les suivants, dans l'ordre de leur fréquence dans le corpus :

1.	Registrul Electronic Național	213
2.	Vaccinarea împotriva COVID-19	205
3.	aplicația Registrul Electronic Național	203
4.	evidența persoanelor vaccinate	159
5.	coronavirus	150
6.	campanie de vaccinare	130
7.	tranșă de vaccin	114
8.	prima doză	90
9.	vaccinare	90
10.	al patrulea val	87
11.	centre de vaccinare	79
12.	campania de vaccinare	71
13.	centrele de vaccinare	70
14.	reacții adverse	69
15.	doze de vaccin	69
16.	doză de rapel	67
17.	vaccinare împotriva COVID-19	64
18.	vaccin Pfizer	45
19.	doză de vaccin	41
20.	val al pandemiei	40
21.	prima doză	40
22.	tranșă	39
23.	a doua doză	38
24.	centre de stocare	37
25.	vaccinul Pfizer BioNTech	37
26.	zone în carantină	36
27.	persoane infectate	36
28.	persoane vaccinate	36
29.	patrulea val al	36
30.	campanie de vaccinare	34
31.	doza de rapel	33
32.	vaccinate împotriva COVID-19	32
33.	caz de infectare	32
34.	medic de familie	32
35.	situație de urgență	32
36.	tranșă de vaccin	32

Comme dans le cas des termes monolexicaux, on constate que la plupart des termes font référence à la prévention de l'infection avec la Covid-19 et que 27 des termes identifiés et extraits appartiennent au champ lexical *vaccinare.* Avec 27 termes sur les 36 qui composent la liste totale que nous avons analysée,

on peut affirmer que *vaccinarea* représente 75 % de la terminologie la plus fréquemment utilisée dans le corpus. De manière surprenante, les termes polylexicaux les plus utilisés dans la communication officielle et les médias en Roumanie sont les suivants *Registrul Electronic Național,* avec 213 occurrences, *Vaccinarea împotriva COVID-19*, avec 205 occurrences et *aplicația Registrul Electronic Național*, avec 203 occurrences dans le corpus.

Ces résultats nous permettent d'affirmer que dans la communication au sujet de la pandémie en Roumanie, on a notamment mis l'accent sur l'action de vaccination et sur l'enregistrement des personnes vaccinées, d'abord dans le Registre électronique national, puis dans l'application qui génère le Passeport sanitaire européen.

4.3. Mots-clés

Dans le but de disposer d'une source complémentaire pour confirmer l'exactitude et la fiabilité des résultats de l'analyse du corpus, il a été décidé de générer une liste de mots-clés en utilisant la fonction *Keywords* de la plateforme *SketchEngine*. Tout comme pour la liste des termes polylexicaux, une liste de mots-clés a été générée automatiquement. Même si la liste est apparemment similaire, pour obtenir les mots-clés, une comparaison est faite entre le corpus analysé et un corpus de référence en roumain appartenant à la plateforme *SketchEngine*.

La liste ainsi obtenue est similaire mais non identique à la précédente et étant toutes deux basées sur la fréquence d'utilisation, même si les deux listes sont différentes, elles devraient générer des résultats comparables, confirmant ainsi les termes les plus fréquemment utilisés dans la communication officielle et les médias en roumain mentionnés dans la liste précédente. La même procédure a été suivie pour la liste des mots-clés, les 500 premières occurrences de la liste générée étant examinées, ce qui a permis d'identifier 50 mots-clés, classés en fonction de leur fréquence d'utilisation.

1. centru de vaccinare
2. vaccinare împotriva covid-19
3. doză de vaccin
4. campanie de vaccinare
5. registrul electronic național
6. vaccinat împotriva covid-19
7. tranșă de vaccin
8. al patru val
9. doză de rapel
10. vaccin împotriva covid-19
11. caz de covid-19
12. noua tranșă de vaccin
13. reacție adversă
26. prima doză pfizer
27. doză de vaccin
28. cu prima doză pfizer
29. val al pandemiei
30. apă și săpun
31. maraton de vaccinare
32. vaccinare împotriva covid-19 în
33. dozele vor fi depozitate
34. centrele de stocare
35. vaccinul pfizer biontech
36. zonele in carantina din
37. împotriva covid-19
38. transportul către centrele de vaccinare

14. pacient cu covid-19
15. să se vaccineze
16. de reacţii adverse
17. manual pentru prevenirea
18. doze
19. prevenirea şi tratamentul
20. persoane cu prima doză
21. agenţia europeană a medicamentului
22. persoane infectate
23. doza de rapel
24. infectare cu
25. strategia de vaccinare
39. carantina
40. zonele in carantina
41. caz de covid-19
42. al patrulea val
43. val
44. persoane infectate cu virusul
45. persoane vaccinate împotriva covid-19
46. campanie de vaccinare
47. doza de rapel pfizer
48. declarate vindecate şi externate
49. tranşă de vaccin
50. pacient cu covid-19

Ainsi, dans la liste de contrôle des mots-clés, 30 des 50 termes appartiennent au sous-domaine de la *prévention des maladies* et font référence à la vaccination. Par conséquent, dans cette liste de contrôle, 60% des termes polylexicaux obtenus de manière différente font référence à la vaccination. Ces données empiriques nous permettent d'affirmer que, sans aucun doute, dans la période comprise entre mars 2020 et septembre 2021, le sujet le plus discuté dans les documents officiels et les médias en Roumanie a été la vaccination, même si chronologiquement c'est un sujet qui se développe en priorité dans le stade le plus récent du développement du corpus. On voit ainsi comment un sujet récent peut dépasser en visibilité et en l'intérêt du grand public d'autres sujets débattus dans l'espace public bien avant lui.

L'analyse de corpus nous permet également de répondre à la question de la présence de néologismes terminologiques dans le domaine émergent de la Covid-19. Comme on peut le constater à partir des listes de termes monolexicaux, polylexicaux et de mots-clés analysés, la grande majorité du vocabulaire pandémique n'est pas nouvelle. Il s'agit d'un vocabulaire spécialisé existant qui fait partie du vocabulaire commun en très peu de temps et donne l'impression d'un langage néologique car, pendant la pandémie, la façon de parler du grand public, des médias et des autorités locales et nationales change complètement.

En même temps, l'interrogation du corpus nous permet d'affirmer que les néologismes terminologiques de la Covid-19 en roumain sont prédominants des termes polylexicaux et peu présents parmi les termes monolexicaux.

Quant à l'affirmation selon laquelle la majeure partie d'une terminologie émergente est néologique, nous considérons que dans le cas de la Covid-19 en roumain, cette affirmation ne tient pas, car les termes néologiques ne prédominent ni en nombre ni en fréquence d'utilisation dans ce domaine.

5. Conclusions

En conclusion, nous espérons avoir réussi à démontrer que dans le domaine émergent de la Covid-19 en roumain, la terminologie utilisée se compose à la fois de termes existants et de termes néologiques. Malgré le fait que certaines études affirment que dans les terminologies émergentes les néologismes terminologiques prédominent, l'interrogation du *Corpus de textes écrits en roumain dans le domaine Covid-19* ne nous permet pas d'affirmer cela. Il est vrai que de nombreux termes, notamment les termes polylexicaux, sont des néologismes terminologiques, mais on ne peut pas affirmer, sur la base des données empiriques observées à partir de l'analyse du corpus, que les néologismes terminologiques prédominent dans la terminologie de la Covid-19.

Nous pensons également avoir réussi à fournir des arguments quantitatifs et qualitatifs ainsi que des exemples illustratifs issus du corpus pour soutenir l'idée que dans la communication officielle et dans les médias en Roumanie, le principal sujet de débat concernant la Covid-19 est lié à la prévention de la maladie par le moyen de la *vaccination*. Ainsi, tous les résultats obtenus à partir de l'interrogation du corpus soutiennent cette idée tant par la riche variété des termes de la catégorie de la vaccination que par la fréquence élevée d'utilisation qu'ils présentent dans le corpus analysé.

Une autre information importante qui vient d'esquisser le profil de la communication officielle et médiatique en roumain est l'importance accordée en communication aux sous-domaines de la Covid-19. Ainsi, le fait que la prévention soit au premier plan de la communication sur la Covid-19 est cohérent avec les résultats des requêtes ultérieures au niveau des termes sur le corpus. On peut également constater que le degré de pertinence des sous-domaines de la Covid-19 est spécifique à une communication de niveau général et non à une communication entre spécialistes. Par conséquent, on peut affirmer que, pour le grand public, dans la communication quotidienne les sujets les plus importants sont *: mijloacele de prevenţie, aspectele epidemiologice, noţiunile generale et măsurile de protecţie.* Dans un corpus de textes hautement spécialisés, on s'attendrait à ce que des sous-domaines tels que *manifestările clinice*, *diagnosticul* et *tratamentul* soient prioritaires et mieux représentés au niveau de la fréquence d'utilisation.

Bien évidemment, cette analyse peut être approfondie et fournir des résultats encore plus précis en ce qui concerne la répartition des néologismes terminologiques et des termes déjà existants dans le domaine de la Covid-19 en roumain. Néanmoins, pour cela, une comparaison entre les résultats de l'interrogation du corpus et l'analyse qualitative et quantitative du glossaire des termes de la Covid-19 en roumain devra être effectuée. Nous avons l'intention de réaliser cette analyse dans une prochaine étude.

BIBLIOGRAPHIE

Asif, Muhammad, et al. 2021. "Linguistic analysis of *neologism related to coronavirus (COVID-19).*" *Social Sciences & Humanities Open*, 4(1), https://doi.org/10.1016/j.ssaho.2021.100201 (consulté le 10/09/2021).

Humbley, John. 2009. « La terminologie française du commerce électronique, ou comment faire du neuf avec de l'ancien dans Terminologie et plurilinguisme dans l'économie internationale ». Milano : Università Cattolica, 9 juin 2009, https://www.unilat.org/Library/Handlers/File.ashx?id=09850e3c-875c-4fb7-be6c-ae5cc6cdc5e0 (consulté le 10/09/2021).

Kageura, Kyo. 2002. *The dynamics of terminology: A descriptive theory of term formation and terminological growth*. Amsterdam-Philadelphia: John Benjamins Publishing Company.

L'Homme, Marie-Claude. 2004. *La terminologie : principes et techniques*. Montréal : Presses de l'Université de Montréal, http://books.openedition. org/pum/ (consulté le 10/09/2021).

McEnery, Tony, and Andrew Hardie. 2012. *Corpus Linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

Roig-Marín, Amanda. 2020. "English-based coroneologisms. A short survey of our Covid-19-related vocabulary." *English Today*, 1-3, https://doi.org/10.1017/S0266078420000255 (consulté le 10/09/2021).

Rondeau Guy. 1984. *Introduction à la terminologie*, 2e édition. Québec : Éditions Gaëtan Morin.

Samylicheva, Nadezhda, Jiří Gazda. 2020. "Derivative neologisms as sociocultural dominants in the Russian and Czech languages of the modern period." *SHS Web of Conference*s, 88, https://doi.org/10.1051/shsconf/20208801022 (consulté le 10/09/2021).

Sinclair, John. 2005. "Corpus and Text: Basic Principles." In *Developing Linguistic Corpora: a Guide to Good Practice*, edited by Martin Wynne. Oxbow Books https://users.ox.ac.uk/~martinw/dlc/chapter1.htm (consulté le 10/09/2021).

Wynne, Martin (ed.). 2005. *Developing Linguistic Corpora: a Guide to Good Practice*. Oxbow Books, https://users.ox.ac.uk/~martinw/dlc/index.htm (consulté le 10/09/2021).