Un corpus de reformulation des termes du Covid extrait des Archives du Web

Ioana Buhnila, Amalia Todirascu

▶ **To cite this version:**

Ioana Buhnila, Amalia Todirascu. Un corpus de reformulation des termes du Covid extrait des Archives du Web. Humanités numériques en pédagogie et en recherche à la faculté des langues, Sep 2023, Strasbourg, France. Faculté des langues, Université de Strasbourg, 2024, 10.34847/nkl.73c9k3qm . hal-04796991

HAL Id: hal-04796991
https://hal.science/hal-04796991v1

Submitted on 21 Nov 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Faculté des **langues**
Université de Strasbourg

Projet AdaptMed – reformulations des termes du Covid-19

Amalia Todirascu & Ioana Buhnila, LiLPa UR 1339

Objectifs du projet AdaptMed

- **Explorer** les archives du Web pour extraire les textes qui contiennent des reformulations de termes liés au Covid-19 en collaboration avec le **DataLab** de la **BNF** (Bibliothèque nationale de France) et la **BNU** (Bibliothèque nationale universitaire de Strasbourg)
- **Annoter et analyser** les phrases qui contiennent des termes liés au Covid-19 et des marqueurs de reformulation (de type « c'est-à-dire », « est une maladie », « signifie ») pour identifier des reformulations
- Créer un **système de reconnaissance automatique** de reformulations médicales

Motivation

- **Les textes de vulgarisation** expliquent les **termes médicaux** pour le grand public à travers différents types de **reformulations** (paraphrases, définitions, exemplifications, dénominations, explications)
- **La reformulation** - *dire les choses autrement* (*Inkova, 2020)* est une notion complexe en linguistique et, ensemble avec les **termes médicaux**, sont difficiles à identifier par des systèmes de traitement automatique des langues

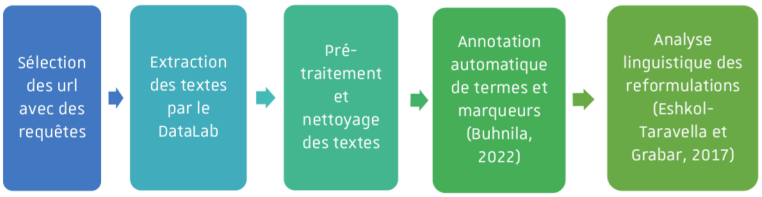
Données : collection Covid-19

- Archive du Web éphémère de la première vague de **l'épidémie Covid-19** février-juillet 2020 (217 millions d'URL, 15 To de données) (accès limité), crée par **DataLab, BNF :**
  - interface graphique avec un langage de requêtes complexe
  - Solr, moteur de recherche experte sans interface graphique

Ressource : dictionnaire DiCovid-19

- **DiCovid-19** : premier dictionnaire de termes utilisés ou apparus pendant la pandémie du Covid-19 (https://dicovid19.com/)
- Exemples de termes covid : *agueusie, cas contact, orage cytokinique, super spreader, vaccinodrome*

Méthodologie



Résultats

Corpus	N° phrases			Annotation phrases		
	Total	Avec termes	Avec termes+ marq	Avec ref	Ref+2	Total refs
BNF1	8995	3340	813	294 (36,16%)	95	389
BNF2	12954	3631	579	182 (31,43%)	56	238
BNF3	3695	1594	333	161 (43,34%)	25	181

- Exemple de requête pour la recherche experte : text: "distanciation physique signifie" ~7 AND (collections:"épidémie Covid-19")
- Exemples de reformulations : **distanciation physique** d'autres que cela *signifie* couper les contacts sociaux ;
  l'anosmie, *c'est-à-dire* une perte totale de l'odorat.

Statut	Phrase avec terme et marqueur	Terme	Marqueur	Reformulation	Relations lexicales	Fonctions sémantico-pragmatiques
oui	Un nouveau variant du Covid-19, appelé 501.V2, est apparu en octobre en Afrique du Sud.	nouveau variant du Covid-19	appelé	501.V2	synonymie	dénomination

Conclusions et perspectives

- **868** termes médicaux liés à la Covid-19 et leurs reformulations ont été annotés et analysés en termes de relations lexicales et fonctions sémantico-pragmatiques
- Développement d'une **méthode de construction de requêtes** d'identification des reformulations médicales
- Les données seront intégrées dans un **système automatique de génération des reformulations médicales**

Références

Buhnila Ioana. (2022). Le Rôle Des Marqueurs et Indicateurs Dans l'analyse Lexicale et Sémantico-Pragmatique de Reformulations Médicales. 8e Congrès Mondial de Linguistique Française(CMLF), 4-8 juillet 2022, Orléans, France, SHS Web of Conferences 138: 10005. https://doi.org/10.1051/shsconf/202213810005.

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

Eshkol-Taravella Iris & Grabar Natalia. (2017). Taxinomie dans les reformulations du point de vue de la linguistique de corpus. Syntaxe et Sémantique, vol. 18, no. 1, pp. 149-184.

Inkova Olga. (dir.). (2020). Autour de la reformulation. Droz, collection Recherches et rencontres, Publication de la Faculté des Lettres de l'Université de Genève, no. 36, 216 pages.

todiras@unistra.fr ioana.buhnila@etu.unistra.fr