

Europresse : Constitution et Exploitation de Corpus

Gaël Lejeune (feat. Thibault Grison)

26 octobre 2023

STIH, Sorbonne Université

Problématique de la constitution de corpus

1. Constituer des corpus **homogènes, représentatifs**
2. satisfaisant des critères de **qualité**
3. **utilisables** avec des outils informatiques

1. « Construire et utiliser un corpus : le point de vue d'une sémantique textuelle interprétative », Atelier Corpus et TAL : pour une réflexion méthodologique

Problématique de la constitution de corpus

1. Constituer des corpus **homogènes, représentatifs**
2. satisfaisant des critères de **qualité**
3. **utilisables** avec des outils informatiques

Ou de façon plus formelle (Pincemin 1999¹) :

Signifiante : cohérence (objet/angle déterminé), pertinence (étude déterminée)

Acceptabilité : représentativité (fidèle), régularité (non-parasité), complétude (ampleur)

Exploitabilité : homogénéité (commensurable), volume (significatif)

1. « Construire et utiliser un corpus : le point de vue d'une sémantique textuelle interprétative », Atelier Corpus et TAL : pour une réflexion méthodologique

Table 1 – Quelques solutions existantes (Presse fr)

Corpus	Homog.	Qual.	Util.	Taille
Est républicain (Ortolang)	Partiel	+++	+	++
Web As Corpus (Sketch Engine)	Tout venant	+	++	+++
Google News	OK	++	--	++
Europresse	++	+++	???	+++

- Moteur de recherche OK mais :
 - Comment exploiter dans un concordancier ?
 - Conserver les corpus pour des exploitations futures ?

- Moteur de recherche OK mais :
 - Comment exploiter dans un concordancier ?
 - Conserver les corpus pour des exploitations futures ?

Limites rencontrées :

1. Résultats par tranche de 100 documents (limite de 1 000)
2. format peu adapté à la requête plein texte (PDF)
3. Utilisabilité des méta-données ,

- Moteur de recherche OK mais :
 - Comment exploiter dans un concordancier ?
 - Conserver les corpus pour des exploitations futures ?

Limites rencontrées :

1. Résultats par tranche de 100 documents (limite de 1 000)
2. format peu adapté à la requête plein texte (PDF)
3. Utilisabilité des méta-données ,

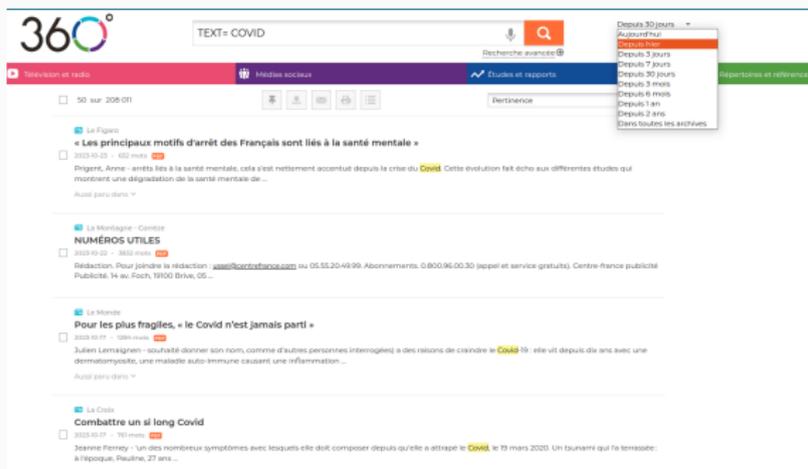
Solutions :

1. Du bricolage : une recherche temporelle pas pratique à manipuler
2. Du sioutage : exports + version "classique" d'Europresse
3. Du détournage (web parsing) : l'outil `EUROPARSER`

1 : obtenir du corpus (I)

Bricolage

Un objet lourd (mais non contondant) et la touche page down (mais on reste limité à 1 000 résultats)



The screenshot displays a search interface on the 360° website. The search bar contains the text "TEXT= COVID". Below the search bar, there are navigation tabs for "Télévision et radio", "Médias sociaux", "Études et rapports", and "Portables et références". A dropdown menu is open, showing time-based filters: "Depuis 30 jours", "Aujourd'hui", "Depuis 1 semaine", "Depuis 5 jours", "Depuis 7 jours", "Depuis 30 jours", "Depuis 3 mois", "Depuis 6 mois", "Depuis 1 an", "Depuis 2 ans", and "Statut toutes les archives". The search results list includes:

- La France**: « Les principaux motifs d'arrêt des Français sont liés à la santé mentale » (2023-10-25 - 652 mots)
- La Montagne - Centre**: NUMÉROS UTILES (2023-10-22 - 3610 mots)
- Le Monde**: Pour les plus fragiles, « le Covid n'est jamais parti » (2023-10-17 - 108 mots)
- La Croix**: Combattre un si long Covid (2023-03-17 - 168 mots)

Figure 1 – Version de base : résultats

1 : obtenir du corpus (II)

The screenshot shows a web interface with a top navigation bar containing three sections: 'Télévision et radio' (red), 'Médias sociaux' (purple), and 'Études et rapports' (blue). Below the navigation bar, there is a search bar with the text '50 sur 208 313' and a dropdown menu set to 'Pertinence'. A toolbar with icons for search, download, share, and list is visible. A context menu is open over the first search result, offering two options: 'Exporter au format PDF' and 'Exporter au format RIS'. The search results list several articles:

- Le Figaro**: « Les principaux motifs d'arrêt des Français en matière de santé mentale » (2023-10-23, 462 mots). Snippet: "Prigent, Anne - arrêts liés à la santé mentale, cela s'est nettement accentué depuis la crise du Covid. Cette évolution fait écho aux différentes études qui montrent une dégradation de la santé mentale de ...".
- Le Montagne - Centre**: NUMÉROS UTILES (2023-10-22, 3832 mots). Snippet: "Rédaction. Pour joindre la rédaction : jussel@centrefrance.com ou 05.55.20.49.99. Abonnements. 0.800.96.00.30 (appel et service gratuits). Centre-france publicité. 14 av. Foch, 19100 Brive, 05 ...".
- Le Monde**: Pour les plus fragiles, « le Covid n'est jamais parti » (2023-10-17, 1294 mots). Snippet: "Julien Lemaignan - souhaité donner son nom, comme d'autres personnes interrogées) à des raisons de craindre le Covid-19 : elle vit depuis dix ans avec une dermatomyosite, une maladie auto-immune causant une inflammation ...".
- La Croix**: Combattre un si long Covid (2023-10-07, 761 mots). Snippet: "Jeanne Ferrey - 'un des nombreux symptômes avec lesquels elle doit composer depuis qu'elle a attrapé le Covid, le 19 mars 2020. Un tsunami qui l'a terrassée: à l'époque, Pauline, 27 ans ...".
- Les Echos**: (no snippet visible).

Figure 2 – Version de base : exports

1 : obtenir du corpus (III)

Sommaire

Documents sauvegardés • 50 documents

Le Figaro	23 octobre 2023 « Les principaux motifs d'arrêt des Français sont liés à la santé mentale » ... arrêts liés à la santé mentale, cela s'est nettement accentué depuis la crise du Covid . Cette évolution fait écho aux différentes études qui montrent une dégradation de la santé mentale de ...	7
La Montagne	22 octobre 2023 NUMÉROS UTILES ... Rédaction. Pour joindre la rédaction : ussel@centrefrance.com ou 05.55.20.49.99. Abonnements. 0.800.96.00.30 (appel et service gratuits). Centre-france publicité Publicité, 14 av. Foch, 19100 Brive, 05 ...	9
Le Monde	17 octobre 2023 Pour les plus fragiles, « le Covid n'est jamais parti » ... souhaité donner son nom, comme d'autres personnes interrogées) a des raisons de craindre le Covid-19 : elle vit depuis dix ans avec une dermatomyosite, une maladie auto-immune causant une inflammation ...	16
La Croix	17 octobre 2023 Combattre un si long Covid ... 'un des nombreux symptômes avec lesquels elle doit composer depuis qu'elle a attrapé le Covid , le 19 mars 2020. Un tsunami qui l'a terrassée : à l'époque, Pauline, 27 ans ...	18
Les Echos	17 octobre 2023 Pfizer et BioNtech peinent à écouler leurs produits Covid Produits Covid cherchent preneurs. Pfizer va devoir en trouver, Outre-Atlantique, pour écouler les 7,9 millions de traitements Paxlovid dont l'Etat fédéral américain ne veut plus. A la suite d ...	20

Figure 3 – Version de base : visualisation

1 : obtenir du corpus (IV)

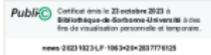
Documents sauvegardés par Sorbonne Université

Lundi 23 octobre 2023 à 12 h 02

Documents sauvegardés

LE FIGARO

© 2023 Le Figaro. Tous droits réservés.
Le présent document est protégé par les lois et conventions internationales sur le droit d'auteur et son utilisation est régie par ces lois et conventions.



Nom de la source	Lundi 23 octobre 2023
Le Figaro	La Figaro • no. 24624
Type de source	• p. 13
Presse • Journaux	• 652 mots
Périodicité	Quotidien
Couverture géographique	Santé
Nationale	
Provenance	
France	



« Les principaux motifs d'arrêt des Français sont liés à la santé mentale »

Prigent, Anne

RESPONSABLE de la Mission animation du service médical à la Caisse nationale de l'assurance-maladie, le Dr Rémi Pécault-Charby fait le point sur les causes de souffrance au travail, et les actions mises en oeuvre.

LE FIGARO. - Quelles sont les principales causes médicales d'arrêt de travail et celles en progression ?

Dr Rémi PÉCAULT-CHARBY. - En 2022, les principaux motifs d'arrêt de travail des Français sont liés à la santé

représentent environ 7,5 % des arrêts. Pour les arrêts de plus de six mois, on retrouve principalement les syndromes dépressifs et troubles anxieux dépressifs et les lombalgies chroniques.

Que représentent les Covid longs dans les arrêts de travail ?

Nous avons des difficultés pour tracer spécifiquement les personnes atteintes de Covid long. En effet, le tableau clinique peut être assez polymorphe et on peut avoir des patients en affection longue durée pour Covid long ou pour asthénie, voire pour des troubles liés à

Pour moitié, cette hausse est liée à des facteurs mécaniques : volume d'activité de la population active en augmentation, vieillissement de la population active ou encore augmentation des salaires à partir desquels sont calculées les indemnités journalières. L'autre moitié est due à une augmentation du recours aux arrêts de travail ainsi qu'à l'allongement de leur durée. Ce sont des tendances de fond sur lesquelles on peut s'interroger : est-ce lié à la dégradation générale de la santé mentale de la population, à une hausse des risques psychosociaux (c'est-à-dire des éléments qui portent atteinte à l'intégrité physique et à la santé men-

Figure 4 – Version de base : visualisation

2 : obtenir un vrai corpus grâce à la version "classique" (I)

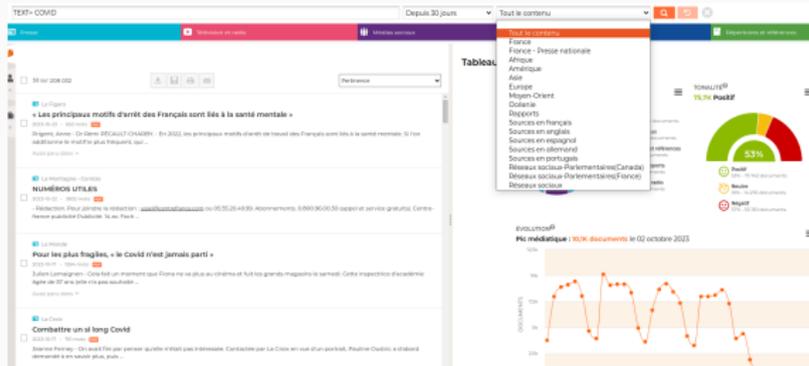


Figure 5 – Version "classique"

2 : obtenir un vrai corpus grâce à la version "classique" (II)

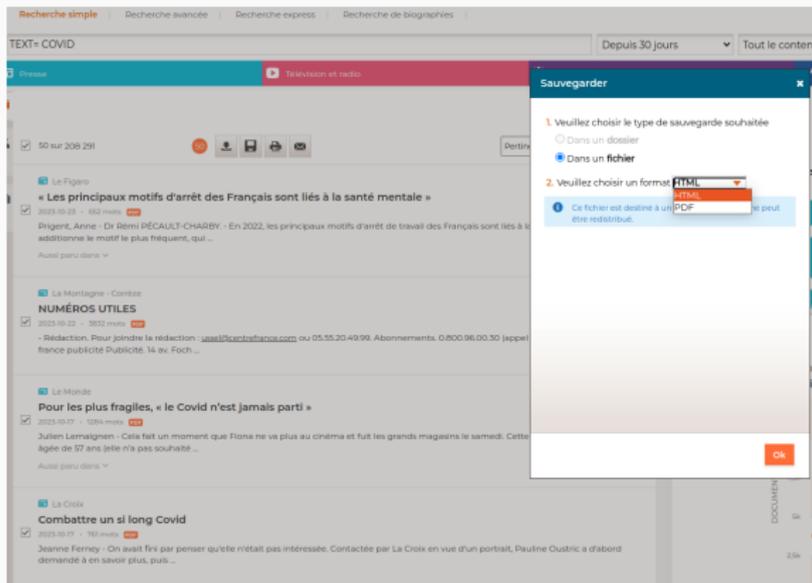
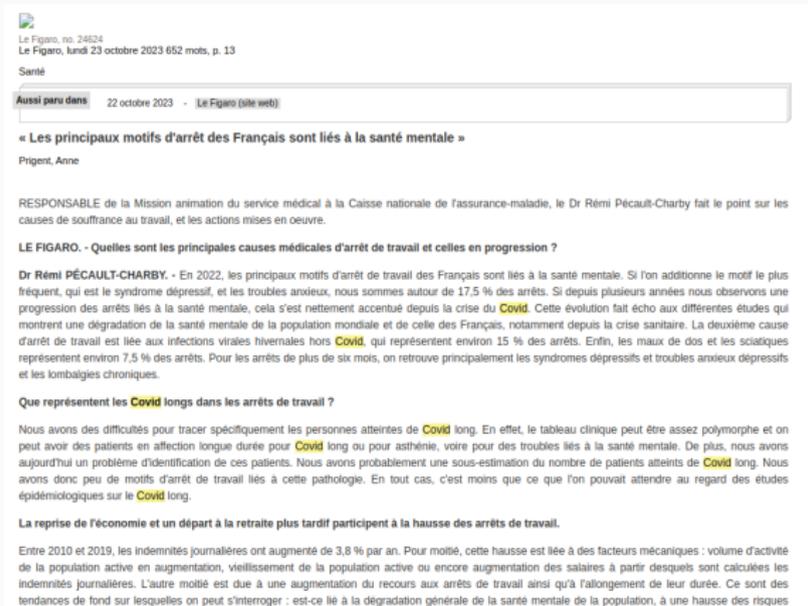


Figure 6 – Version de base : exports

2 : obtenir un vrai corpus grâce à la version "classique" (III)



Le Figaro, no. 24624
Le Figaro, lundi 23 octobre 2023 652 mots, p. 13

Santé

Aussi paru dans 22 octobre 2023 - [Le Figaro \(site web\)](#)

« Les principaux motifs d'arrêt des Français sont liés à la santé mentale »

Prigent, Anne

RESPONSABLE de la Mission animation du service médical à la Caisse nationale de l'assurance-maladie, le Dr Rémi Pécault-Charby fait le point sur les causes de souffrance au travail, et les actions mises en oeuvre.

LE FIGARO. - Quelles sont les principales causes médicales d'arrêt de travail et celles en progression ?

Dr Rémi PÉCAULT-CHARBY. - En 2022, les principaux motifs d'arrêt de travail des Français sont liés à la santé mentale. Si l'on additionne le motif le plus fréquent, qui est le syndrome dépressif, et les troubles anxieux, nous sommes autour de 17,5 % des arrêts. Si depuis plusieurs années nous observons une progression des arrêts liés à la santé mentale, cela s'est nettement accentué depuis la crise du **Covid**. Cette évolution fait écho aux différentes études qui montrent une dégradation de la santé mentale de la population mondiale et de celle des Français, notamment depuis la crise sanitaire. La deuxième cause d'arrêt de travail est liée aux infections virales hivernales hors **Covid**, qui représentent environ 15 % des arrêts. Enfin, les maux de dos et les sciatiques représentent environ 7,5 % des arrêts. Pour les arrêts de plus de six mois, on retrouve principalement les syndromes dépressifs et troubles anxieux dépressifs et les lombalgies chroniques.

Que représentent les **Covid longs dans les arrêts de travail ?**

Nous avons des difficultés pour tracer spécifiquement les personnes atteintes de **Covid** long. En effet, le tableau clinique peut être assez polymorphe et on peut avoir des patients en affection longue durée pour **Covid** long ou pour asthénie, voire pour des troubles liés à la santé mentale. De plus, nous avons aujourd'hui un problème d'identification de ces patients. Nous avons probablement une sous-estimation du nombre de patients atteints de **Covid** long. Nous avons donc peu de motifs d'arrêt de travail liés à cette pathologie. En tout cas, c'est moins que ce que l'on pouvait attendre au regard des études épidémiologiques sur le **Covid** long.

La reprise de l'économie et un départ à la retraite plus tardif participent à la hausse des arrêts de travail.

Entre 2010 et 2019, les indemnités journalières ont augmenté de 3,8 % par an. Pour moitié, cette hausse est liée à des facteurs mécaniques : volume d'activité de la population active en augmentation, vieillissement de la population active ou encore augmentation des salaires à partir desquels sont calculées les indemnités journalières. L'autre moitié est due à une augmentation du recours aux arrêts de travail ainsi qu'à l'allongement de leur durée. Ce sont des tendances de fond sur lesquelles on peut s'interroger : est-ce lié à la dégradation générale de la santé mentale de la population, à une hausse des risques

Figure 7 – Version de base : visualisation (moins jolie mais copier-coller plus aisé)

3 : exploiter tout le corpus

Détourage avec EUROPARSER (F. Alié, M.Hernandez, G.Lejeune, S.Poder)

Récupérer les textes, séparés et structurés pour pouvoir affiner l'analyse avec :

- IRAMUTEQ
- EXCEL
- TXM
- ANTCOUC
- VOYANT TOOLS...

Méta-données disponibles : titre, source, date, (auteur)

Dé-doublonnage partiel, rangement par dossier ...