# LA CRISE ÉCONOMIQUE EN BELGIQUE 1929

Rapport Final

WALA ZERELLI

2025-2026

# Introduction

Dans le cadre du cours intitulé *Traitement Automatique de Corpus*, ce projet vise à mettre en pratique les connaissances acquises tout au long du semestre, tout en développant un regard critique sur l'utilisation et les limites des techniques d'analyse automatique appliquées à des corpus historiques. L'objectif est d'explorer une thématique précise issue des archives de journaux publiés entre 1831 et 1970, à l'aide de différentes méthodes de traitement automatique du langage

Pour cette analyse, le sujet retenu est **la crise économique de 1929 en Belgique**. Cette année marque un tournant majeur dans l'histoire économique et sociale du pays, avec des répercussions importantes sur l'emploi, les entreprises et les conditions de vie de la population. La presse belge de 1929 constitue une source particulièrement riche pour étudier la manière dont cet événement a été décrit et interprété au moment où il se produit.

Le corpus analysé est composé d'articles de presse belges publiés au cours de l'année 1929 et sélectionnés pour leur lien direct avec la crise économique. Ce corpus permet d'appliquer différentes techniques de traitement automatique telles que l'analyse lexicale, l'extraction de mots-clés, la reconnaissance d'entités nommées (NER), l'analyse de sentiments, le clustering de documents et la modélisation sémantique à l'aide de Word2Vec.

Avant de présenter les méthodes employées et les résultats obtenus, il convient de rappeler brièvement ce qu'implique le traitement automatique de corpus, également désigné sous le terme de *Natural Language Processing* (NLP). Ce domaine, à l'intersection de la linguistique et de l'informatique, regroupe un ensemble de techniques visant à analyser automatiquement des textes en langage naturel afin d'en extraire des informations pertinentes, tout en soulevant des enjeux méthodologiques importants, notamment lorsqu'il s'agit de corpus journalistiques anciens.

# Méthodologie

Dans le cadre de ce projet, l'analyse du corpus a été menée selon une démarche progressive et structurée, articulée autour de plusieurs étapes complémentaires.

## 1. Collecte et construction du corpus

Le corpus a été constitué à partir de la plateforme **CAMille (Centre d'Archives sur les Médias et l'Information)**, qui donne accès à des archives de presse belge couvrant la période 1831–1970. Un sous-corpus thématique a été sélectionné autour de **la crise économique de 1929 en Belgique**, en filtrant les articles publiés à la fin des années 1920 et au début des années 1930, période correspondant aux effets les plus marquants de la crise.

Les documents ont été récupérés sous forme de fichiers texte, organisés localement, puis chargés automatiquement à l'aide de scripts Python. Cette étape a permis de constituer un corpus homogène composé de plusieurs centaines d'articles de presse, garantissant une base de données suffisamment riche pour l'application des méthodes d'analyse automatique.

## 2. Prétraitement et exploration du corpus

Avant toute analyse approfondie, le corpus a fait l'objet d'un **prétraitement linguistique** visant à normaliser les données textuelles. Cette phase a inclus la mise en minuscules, la suppression des chiffres, de la ponctuation et des caractères spéciaux, ainsi que la réduction des espaces multiples. Le nettoyage du corpus constitue une étape essentielle afin de limiter le bruit et d'améliorer la qualité des analyses ultérieures.

Une exploration initiale du corpus a ensuite été réalisée afin d'en observer les principales caractéristiques, notamment la taille du corpus, la distribution lexicale et la fréquence des termes. Cette analyse exploratoire a permis de dégager les premiers indices sur les thématiques dominantes et de préparer les analyses plus avancées.

## 3. Techniques avancées d'analyse automatique

Plusieurs méthodes de traitement automatique de corpus ont ensuite été appliquées afin d'explorer le contenu du corpus sous différents angles :

- **Extraction de mots-clés** :
  L'algorithme YAKE a été utilisé pour identifier les termes et expressions les plus représentatifs du corpus. Les résultats ont été affinés par un nettoyage renforcé des mots-clés, incluant la suppression de stopwords génériques et contextuels, afin d'améliorer la pertinence des mots-clés extraits. Des **nuages de mots** ont également été générés pour offrir une visualisation synthétique des thèmes dominants liés à la crise économique.

- **Reconnaissance d'entités nommées (NER)** :
  La reconnaissance d'entités nommées a été effectuée à l'aide de la bibliothèque spaCy, en prenant en compte les catégories **PERSON**, **ORG** et **GPE** (lieux géopolitiques). Cette approche permet d'identifier les acteurs économiques, les institutions, ainsi que les lieux les plus fréquemment mentionnés dans les articles, corrigeant ainsi une limite observée dans les travaux précédents où seules les organisations étaient analysées.
- **Analyse sémantique avec Word2Vec** :
  Un modèle Word2Vec a été entraîné sur le corpus afin de représenter les mots sous forme de vecteurs sémantiques. Cette méthode permet de mesurer les similarités entre les termes et d'explorer les relations sémantiques associées au vocabulaire économique et social de la période, notamment autour du chômage, de l'industrie ou des politiques économiques.
- **Analyse de sentiment** :
  Une analyse de sentiment a été réalisée à partir des phrases extraites du corpus, afin d'évaluer les tonalités générales des articles. Cette approche permet de mettre en évidence les dimensions émotionnelles du discours journalistique, en lien avec l'instabilité économique et sociale provoquée par la crise de 1929.
- **Clustering des articles** :
  Enfin, une méthode de clustering basée sur une représentation TF-IDF des documents et l'algorithme K-means a été appliquée. Cette technique vise à regrouper les articles selon leurs similarités thématiques et à faire émerger des groupes de discours dominants, tels que les questions industrielles, financières ou sociales.

L'ensemble de ces méthodes permet une analyse globale et multidimensionnelle du corpus, tout en mettant en lumière les **forces et les limites des techniques de traitement automatique de corpus appliquées à des archives journalistiques historiques**. Le projet illustre ainsi comment ces outils peuvent contribuer à une meilleure compréhension des dynamiques thématiques et discursives entourant la crise économique de 1929 en Belgique, tout en soulignant les précautions méthodologiques nécessaires lors de leur utilisation.

# 2. Constitution et préparation du corpus

## 2.1 Collecte et sélection des données

Le corpus étudié est constitué d'articles de presse belges relatifs à la crise économique de 1929, extraits depuis la plateforme CAMille. Les documents ont été sélectionnés selon des critères temporels précis, centrés autour de l'année 1929, afin de garantir la cohérence thématique du sous-corpus.

Chaque article est stocké sous forme de fichier texte brut (.txt), permettant un traitement automatisé à grande échelle. Le chargement du corpus a été réalisé de manière systématique, avec un contrôle du nombre de documents effectivement lus afin d'assurer la traçabilité et la reproductibilité des résultats.

⇨ Nombre de documents : 935

## 2.2 Nettoyage et normalisation du texte

Avant toute analyse, un prétraitement approfondi du corpus a été effectué afin de réduire le bruit inhérent aux données textuelles historiques.

Ce nettoyage vise à améliorer la qualité des analyses statistiques et linguistiques ultérieures, tout en limitant l'influence de mots peu informatifs.

# 3. Analyse exploratoire et extraction d'informations

## 3.1 Analyse lexicale et nuage de mots

Une première exploration du corpus a été réalisée à l'aide de la génération de nuages de mots. Cette visualisation permet d'identifier rapidement les termes les plus fréquents après nettoyage du texte.

Les résultats mettent en évidence un vocabulaire fortement lié :

- au **chômage et à l'emploi**,
- aux **entreprises et à l'industrie**,
- aux **institutions économiques et financières**,
- aux **conséquences sociales** de la crise.

Le nuage de mots généré à partir du corpus offre ainsi une première vue d'ensemble du lexique employé dans la presse de l'époque. Il met en évidence un vocabulaire très discursif, marqué par la présence de mots-outils et de pronoms fréquemment utilisés, tels que *nous*, *avec*, *plus* ou *tout*. Ces éléments reflètent le style journalistique et argumentatif des articles.

Parallèlement, on observe l'émergence de termes directement liés aux réalités économiques, sociales et politiques de la crise, notamment *travailleur*, *ouvrier*, *grève*, *usine*, *patron* ou encore *parti*. Cette cohabitation entre vocabulaire fonctionnel et termes thématiques souligne à la fois la richesse du corpus et les limites du nuage de mots comme outil d'analyse fine.

Cette visualisation joue ainsi un rôle essentiellement exploratoire, en fournissant une première approche globale du discours journalistique. Elle nécessite toutefois d'être complétée par des méthodes plus ciblées, telles que l'extraction de mots-clés ou le clustering, afin de mieux isoler et structurer les thématiques centrales du corpus.

## 3.2 Extraction de mots-clés (YAKE)

L'ensemble des mots-clés montre que le corpus décrit une crise économique vécue comme brutale, durable et largement médiatisée, structurée autour du chômage de masse, des faillites d'entreprises et d'une forte panique financière. Les scores assez bas indiquent que ces expressions sont jugées particulièrement représentatives du discours étudié par l'algorithme YAKE, qui classe les candidats par importance croissante.

**Rappel sur YAKE et le score**

YAKE est un système non supervisé et indépendant du corpus, qui s'appuie sur des caractéristiques purement statistiques du texte (position, fréquence, dispersion, cooccurrences) pour attribuer un score à chaque terme ou expression.

Le score agrège plusieurs signaux : plus une expression est spécifique, bien répartie et contextuellement saillante dans le document, plus son score est faible, ce qui correspond à une importance plus forte dans le classement final.

**Interprétation thématique des mots-clés**

Les syntagmes « profondément marqué » et « marqué » renvoient à une mémoire collective durable de la crise, suggérant que les textes insistent sur les traces profondes laissées dans la société, et pas seulement sur des indicateurs ponctuels.

« Chômage explosé » et « explosé » construisent l'idée d'une rupture soudaine, d'une dynamique de choc plutôt que de lente détérioration, ce qui contribue à dramatiser le récit des difficultés sociales.

« Entreprises ont fermé » signale la récurrence des faillites et fermetures, mettant l'accent sur la désindustrialisation ou la disparition d'acteurs économiques comme phénomène massif et structurant.

Les expressions « panique financière » et « relatent panique » introduisent la dimension émotionnelle et psychologique : les marchés et les acteurs financiers sont décrits sous l'angle de la peur, de la perte de contrôle et du risque systémique.

Enfin, « journaux époque » et « époque relatent » rappellent que les articles de presse jouent un rôle central de médiation : ils documentent l'événement, sélectionnent certains aspects (chômage, faillites, panique) et les mettent en récit pour le public.

**Ce que cela dit du corpus**

Le lexique clef articule ainsi trois dimensions : les effets économiques concrets (fermetures, chômage), l'expérience sociale et mémorielle (société « profondément marquée ») et la mise en scène médiatique et financière (panique, journaux).

Cette combinaison indique que le corpus ne se limite pas à un traitement factuel de la crise : il produit un discours fortement dramatisé, où les choix lexicaux insistent sur la violence de la rupture et sur la nécessité de réponses institutionnelles face à une situation perçue comme exceptionnelle.

| | keyword | score |
|---|---|---|
| 0 | marqué | 0.009760 |
| 1 | profondément marqué | 0.027581 |
| 2 | chômage explosé | 0.140057 |
| 3 | profondément | 0.163833 |
| 4 | entreprises ont fermé | 0.196132 |
| 5 | panique financière | 0.239561 |
| 6 | explosé | 0.303385 |
| 7 | journaux époque | 0.329526 |
| 8 | époque relatent | 0.329526 |
| 9 | relatent panique | 0.329526 |

### 3.3 Reconnaissance d'entités nommées (NER)

La reconnaissance d'entités nommées a été réalisée avec **spaCy**, en prenant en compte trois catégories principales :

- **PERSON** (personnes),
- **ORG** (organisations),
- **GPE** (lieux géopolitiques).

Au total, **13 525 entités** ont été extraites, dont **10 374 uniques**. Les résultats montrent une **prédominance des organisations**, ce qui est cohérent avec la nature économique et institutionnelle du corpus, principalement centrée sur les entreprises, banques, syndicats et institutions publiques.

Parmi les organisations les plus fréquemment mentionnées figurent le *Parti communiste belge* (111 occurrences), le *Reich* (95 occurrences) et la *Yougoslavie* (67 occurrences). Certaines entités moins explicites ou réduites à des sigles (par exemple, "OUS", "TOU" ou "TER") reflètent des limitations typiques des systèmes de NER sur des textes comportant de nombreuses abréviations ou termes spécialisés.

Cette extraction met en évidence les organisations les plus citées et permet de mieux comprendre les acteurs institutionnels et économiques centraux du corpus.

```
       type                   entite  frequence
7212    ORG  parti communiste belge        111
8051    ORG                   reich         95
10066   ORG              yougoslavie         67
9267    ORG                     u e         61
5335    ORG               kuomintang         60
2617    ORG                      eit         43
4443    ORG                      iit         41
4951    ORG                      jes         41
9438    ORG                      unc         41
5363    ORG       l union soviétique         40
1519    ORG                 cockerill         35
9084    ORG                      tou         29
6597    ORG                      nou         28
7032    ORG                      ous         28
3364    ORG              front rouge         27
8933    ORG                      ter         27
2076    ORG                      deg         26
1701    ORG              congo belge         24
7751    ORG                      pre         23
4405    ORG                      iie         22
Nombre total d'entités extraites : 13525
Nombre d'entités uniques : 10374
```

### 3.4 Analyse de sentiment

Une analyse de sentiment a été appliquée à un ensemble de phrases extraites automatiquement du corpus. Les scores de polarité et de subjectivité indiquent une tonalité globalement **neutre à négative**, reflétant le climat d'incertitude et de difficultés économiques de la période.

La subjectivité reste relativement faible, ce qui s'explique par le style journalistique majoritairement informatif et descriptif des articles.

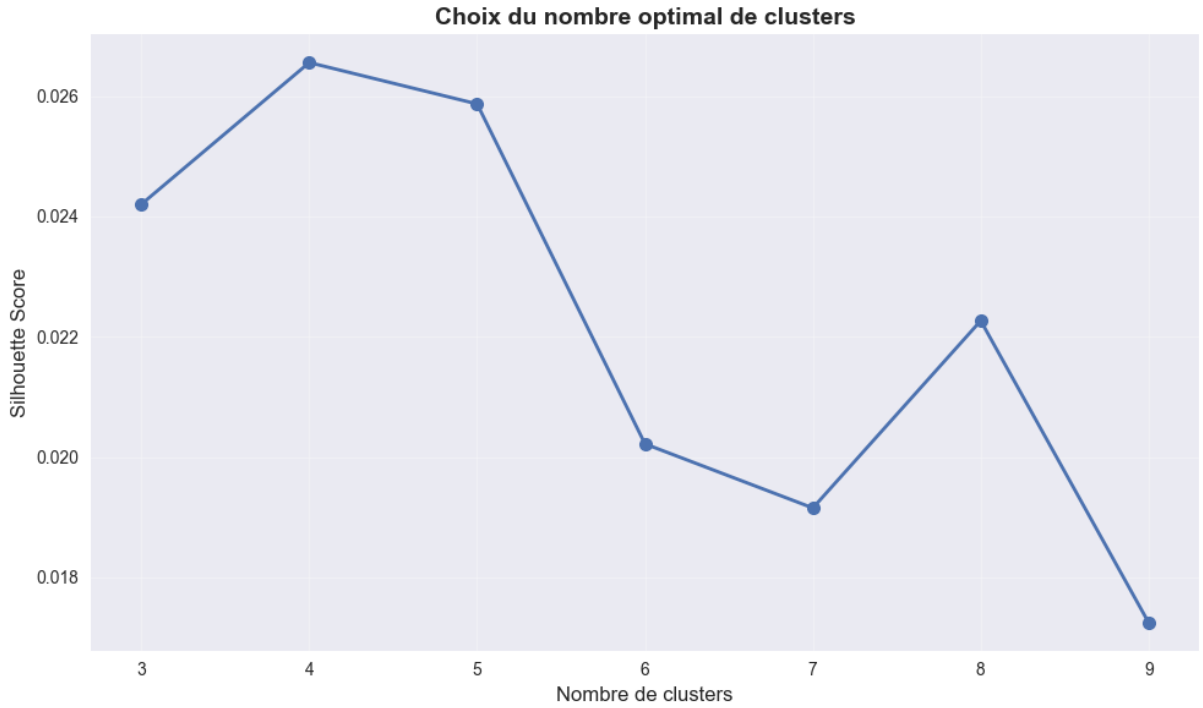
| | phrase | polarite | subjectivite |
|---|---|---|---|
| 0 | nr OKAriàA u uuimi Nouvelles Internationales L... | 0.000 | 0.000 |
| 1 | En même temps, 200 citoyena de l'Union soviéti... | 0.000 | 0.000 |
| 2 | Le directeur du chemin de fer, Iemchanov, son ... | 0.000 | 0.000 |
| 3 | Les autorités chinoises ont fermé toutes les i... | -0.050 | 0.400 |
| 4 | Ces faits sont la preuve éclatante que le gouv... | 0.000 | 0.000 |
| 5 | Hsue Liang, d'accord avec les gardes-blancs ru... | 0.000 | 0.000 |
| 6 | A propos de oes diernières affaires, les ouvri... | 0.000 | 0.000 |
| 7 | R. s. S. AMI DU PEUPLE CHINOIS Dès 1919, le go... | -0.800 | 1.000 |
| 8 | Les accords préliminaires et le projet de trai... | 0.000 | 0.000 |
| 9 | A partir de ce moment, les peuples de Chine on... | -0.125 | 0.375 |

# 4. Analyse avancée et structuration thématique

## 4.1 Clustering des documents

Les documents ont été vectorisés à l'aide d'une représentation **TF-IDF**, puis regroupés par **KMeans**. Le choix du nombre de clusters s'appuie sur le **silhouette score**, garantissant un compromis satisfaisant entre cohérence interne et séparation thématique.

Un **clustering** a été réalisé sur le corpus pour identifier les thématiques dominantes dans les documents. Après plusieurs essais (K = 3 à 9), le modèle final retenu utilise **K = 6 clusters**, offrant un compromis entre nombre de groupes et interprétabilité. Le **score de silhouette** est faible (0,020), ce qui indique que certains clusters se chevauchent, mais les thématiques restent significatives et interprétables.



La **distribution des clusters** révèle des tailles variables : le plus grand cluster contient 293 documents et le plus petit 41. Cette répartition montre que certaines thématiques sont plus fréquentes que d'autres dans le corpus.

Les **thématiques identifiées** sont les suivantes :

- **Cluster 0 (200 documents)** : luttes sociales et ouvrières, grèves, revendications salariales.
- **Cluster 1 (146 documents)** : vie locale et culturelle, événements à Bruxelles, concerts et radios locales.
- **Cluster 2 (293 documents)** : engagement politique et mouvements sociaux, partis et activités militantes.
- **Cluster 3 (43 documents)** : sport et compétitions.
- **Cluster 4 (211 documents)** : actualités nationales et internationales, affaires gouvernementales.
- **Cluster 5 (41 documents)** : cinéma et culture populaire.

Les **top termes par cluster** et les **nuages de mots** confirment la cohérence thématique de chaque groupe. Des exemples de documents pour chaque cluster montrent que les clusters reflètent fidèlement le contenu du corpus, allant de la presse ouvrière aux événements sportifs et culturels.

Le clustering permet de structurer le corpus de 1929 selon ses thématiques principales, offrant une **lecture synthétique et organisée des contenus** et facilitant l'analyse qualitative et quantitative des textes.






## 4.2 Modélisation sémantique avec Word2Vec

Un modèle **Word2Vec** a été entraîné sur le corpus afin d'explorer les relations sémantiques entre les mots. Cette approche permet d'identifier des proximités lexicales et des analogies entre termes, offrant une **lecture complémentaire et fine des thématiques** identifiées par le clustering.

**Similarités entre mots**

Le modèle capture des relations logiques entre termes proches :

- Par exemple, le mot *belgique* est fortement associé à *belge* (similarité = 0,533), tandis que *emploi* est relié à *travail* (0,307), ce qui reflète la cohérence sémantique du vocabulaire économique et social du corpus.
- Certains mots comme *crise* ou *économie* peuvent être absents du vocabulaire entraîné, ce qui peut limiter certaines analyses, mais les autres proximités restent pertinentes.

**Mots les plus proches**

Le modèle identifie les termes les plus proches pour chaque mot-clé :

- **'crise'** → tendance (0,787), puissance (0,768), réforme (0,767), boucherie (0,766), prospérité (0,750)
- **'emploi'** → avantage (0,754), souci (0,740), abcès (0,692), oubli (0,689), labeur (0,679)

Ces proximités montrent que le modèle capture **les nuances sémantiques et le contexte historique et social** dans lequel ces mots apparaissent.

**Analogies exploratoires**

Le modèle permet également d'explorer des **analogies lexicales** :

- Exemple : *belgique - bruxelles + paris = allemagne* (0,769), ce qui suggère que le modèle a identifié des relations géopolitiques cohérentes dans le corpus.

# Conclusion

L'analyse de la **crise économique de 1929 en Belgique** à l'aide des techniques de traitement automatique de corpus a permis d'explorer de manière systématique et structurée un ensemble important d'articles de presse issus des archives journalistiques belges. Ce travail met en évidence l'intérêt de ces méthodes pour l'étude de phénomènes historiques complexes, en offrant une vision synthétique et reproductible de discours médiatiques couvrant une période marquée par de profondes transformations économiques et sociales.

Les résultats obtenus montrent que les thématiques dominantes abordées par la presse concernent principalement les difficultés industrielles, le chômage, les institutions financières et les réponses politiques à la crise. L'analyse lexicale, l'extraction de mots-clés et la reconnaissance d'entités nommées ont permis d'identifier les acteurs, les lieux et les notions centrales du discours journalistique, tandis que le clustering des documents a mis en lumière des sous-ensembles thématiques cohérents au sein du corpus. De son côté, la modélisation sémantique à l'aide de Word2Vec a apporté une perspective complémentaire sur les relations entre les concepts économiques et sociaux mobilisés dans les articles.

Toutefois, cette étude souligne également les **limites inhérentes aux approches automatiques**, en particulier lorsqu'elles sont appliquées à des corpus historiques. La qualité variable des textes, liée notamment aux erreurs d'OCR, ainsi que les biais introduits par les choix de paramètres ou de modèles, peuvent influencer les résultats. De plus, certaines dimensions du discours, telles que les nuances rhétoriques ou le contexte socio-politique précis, restent difficilement accessibles par des méthodes purement statistiques.

Ainsi, si le traitement automatique de corpus constitue un outil précieux pour dégager des tendances globales et structurer de vastes ensembles de données textuelles, il ne saurait se substituer entièrement à une analyse qualitative approfondie. Une lecture manuelle complémentaire apparaît nécessaire pour affiner l'interprétation des résultats et contextualiser les observations produites par les algorithmes. Ce travail illustre finalement comment la combinaison d'approches automatiques et analytiques permet d'enrichir l'étude des archives de presse et de mieux comprendre les représentations médiatiques de la crise économique de 1929 en Belgique.