

TP4 : La « question linguistique » dans une sélection de la presse belge francophone

Guillaume QUINTIN

15 août 2023

Table des matières

1	Introduction	2
2	Choix du sujet et acquisition des données	2
2.1	Bref historique de la « question linguistique » belge	3
2.2	Récupération des textes sur CAMille	4
3	Exploration préliminaire des données	5
4	Mots-clés	6
4.1	Par journal	6
4.2	Focus sur trois périodes	6
5	Nuages de mots	7
6	Reconnaissance d'entités nommées	8
6.1	Par journal	8
6.2	Focus sur trois périodes	8
7	Clustering	9
7.1	Sur le corpus complet	9
7.2	Sur le corpus de chaque journal	10
8	Plongements lexicaux	10
8.1	Création des modèles word2vec	10
8.2	Observations	11
9	Conclusion	13
	Bibliographie	15
10	Annexes	16

1 Introduction

Dans ce travail, nous appliquons des techniques issues du traitement automatique de corpus sur des données de la presse belge francophone. Nous étudierons la façon dont la « question linguistique », qui a animé la vie politique belge pendant une grande partie du XX^e siècle, a été perçue par la presse. Pour ce faire, nous exploiterons les données de trois périodiques francophones qui appartiennent à des tendances idéologiques différentes : La Libre Belgique, Le Peuple et Le Soir. La période concernée commence à la fin de la première guerre mondiale et l’arrivée sur le devant de la scène politique de cette question linguistique, et s’achève en 1950 pour des raisons techniques (nous ne disposons pas des données postérieures pour La Libre Belgique ou pour Le Peuple).

Un des objectifs principaux que nous tenions à poursuivre dans le cadre de ce travail, en plus de vérifier l’efficacité des techniques de traitement automatique sur un corpus historique (ce que nous avons déjà pu réaliser partiellement lors des TP précédents), est de confronter plusieurs journaux entre eux. Ainsi, même si nous aurions pu nous concentrer sur un seul journal mais en englobant une période plus large, nous avons préféré élargir le spectre des périodiques considérés, même si un tel choix entraînait des restrictions au niveau des périodes qui pouvaient être étudiées.

Nous débuterons par poser le décor en synthétisant (trop) brièvement les conflits linguistiques en Belgique. Nous expliquerons également les raisons du choix de notre sujet ainsi que la façon dont nous avons acquis les données relatives à ce dernier. Ensuite, nous effectuerons quelques observations générales sur le corpus (volume, répartition des fichiers...) et appliquerons plusieurs techniques de traitement automatique : l’extraction de mots-clés, la création de nuages de mots, la reconnaissance d’entités nommées, le clustering et les plongements lexicaux. Nous avons décidé de ne pas utiliser l’analyse de sentiments, dont nous ne voyions pas l’intérêt dans le cadre de notre question de recherche. Enfin, nous terminerons en synthétisant les observations réalisées grâce aux différentes techniques mises en œuvre et en portant un regard critique sur l’utilisation de celles-ci sur un corpus de presse historique.

Toutes les figures évoquées dans ce travail sont reproduites dans les annexes, dans leur ordre d’apparition. Des liens dans le texte permettent d’y accéder aisément. Le code Python et les outputs se trouvent dans le fichier `tp4_notebook.ipynb` déposé sur GitHub.

2 Choix du sujet et acquisition des données

Le premier sujet sur lequel nous avons porté notre dévolu était la perception du fédéralisme dans la presse belge. Néanmoins, ce sujet posait quelques difficultés. En effet, il était difficile de sélectionner les documents pertinents : « fédéralisme » fait référence à d’autres réalités que le fédéralisme belge (européen par exemple).

En outre, la presse disponible ne permettait pas une comparaison efficace. La première réforme de l’État date de 1970, et la Belgique ne devient réellement fédérale qu’en 1993, même si la question du fédéralisme est antérieure.¹ Il était nécessaire de considérer uniquement les journaux qui nous sont accessibles jusqu’à la date la plus récente possible. Il s’agit du Soir (jusqu’en 1970) et du Drapeau Rouge (jusqu’en 1966). Toutefois, le volume des données issues du Drapeau Rouge lors de la recherche était bien inférieur à celui du Soir, et cet état de fait ne nous présageait pas une comparaison fructueuse.

1. Étienne ARCQ, Vincent de COOREBYTER et Cédric ISTASSE. « Fédéralisme et confédéralisme ». In : *Dossiers du CRISP* 79.1 (2012), p. 41-44 ; Jean-Claude SCHOLSEM. « Les métamorphoses du fédéralisme ». In : *Bulletin de la Classe des lettres et sciences morales et politiques* 14.1-6 (2003), p. 16-17.

Pour ces raisons, et sans changer entièrement de thématique, nous avons décidé de nous concentrer sur l'une des raisons principales de ce fédéralisme belge relativement récent, à savoir le conflit linguistique entre les deux communautés,² qui agite la vie politique belge depuis la fin du XIX^e siècle et en particulier au siècle suivant.³

2.1 Bref historique de la « question linguistique » belge

À partir de 1830, le français est l'unique langue officielle de la Belgique (en termes de politique, d'administration, d'enseignement, etc.) et la langue de prestige parlée par les élites socio-économiques du pays.⁴ La fin du XIX^e siècle voit une montée en importance du néerlandais, jusqu'à la loi d'égalité de 1898 qui le reconnaît comme langue officielle du pays sur un pied d'égalité avec le français. C'est durant cette période que sont créées, sans les nommer, trois régions administratives linguistiquement distinctes, qui seront à la base de la fédéralisation du pays un siècle plus tard : la Flandre, la Wallonie et Bruxelles.⁵

Après la première guerre mondiale, le conflit linguistique prend le pas sur les conflits confessionnels et religieux qui occupaient la place principale dans la vie politique belge jusqu'alors.⁶ En 1921, la loi Van Cauwelaert reconnaît le néerlandais comme seule langue officielle en Flandre et l'existence désormais ancrée dans la loi de deux régions unilingues (la Flandre et la Wallonie).⁷ Le mouvement flamand, qui s'était construit autour de l'identité linguistique et avait obtenu une première victoire avec la reconnaissance du néerlandais comme langue officielle, revendique également peu à peu une identité politique basée sur cette identité linguistique, tout en prenant de l'ampleur face à la lenteur de la néerlandisation de l'administration en Flandre.⁸ Avec la création officielle de deux régions unilingues, une frontière linguistique est également mise en place, même si concrètement une frontière presque immuable depuis quinze siècles préexistait à son écriture dans la loi. Cette frontière n'a pas bougé malgré l'imposition du français comme langue officielle de la Belgique en 1830. La seule exception est Bruxelles, où le français est présent depuis le XVI^e siècle et dont l'usage n'a cessé de s'étendre au cours du XX^e siècle.⁹ C'est également durant ces années que les cantons germanophones intègrent la Belgique, mais ils n'occuperont jamais une place prépondérante dans le conflit linguistique préexistant.¹⁰

L'état de fait instauré en 1921 est confirmé et renforcé par la loi linguistique de 1932 (loi Jaspar), dans un climat politique tendu en raison d'un bon score des nationalistes flamands aux élections de 1929 et de la néerlandisation de l'université de Gand en 1930. Cette loi promeut un unilinguisme territorial (à l'exception de Bruxelles) et arrête le processus de francisation des grandes villes flamandes. Le principe de territorialité linguistique est également étendu à l'enseignement primaire et moyen, rendant possible pour la première

2. Ce terme est un anachronisme, puisqu'il ne désigne les groupes linguistiques en Belgique que depuis la fédéralisation du pays.

3. Wilfried SWENDEN et Maarten Theo JANS. « 'Will it stay or will it go?' Federalism and the sustainability of Belgium ». In : *West European Politics* 29.5 (2006), p. 878.

4. Claude JAVEAU. « L'énigme de la frontière linguistique en Belgique ». In : *Revue des Sciences Sociales* 48 (2012), p. 49 ; Stéphane RILLAERTS. « La frontière linguistique, 1878-1963 ». In : *Courrier hebdomadaire du CRISP* 2069-2070.24 (2010), p. 8 ; Els WITTE. « La question linguistique en Belgique dans une perspective historique ». In : *Pouvoirs* 136.1 (2011), p. 38.

5. RILLAERTS, « La frontière linguistique, 1878-1963 », p. 9-15.

6. Georges GORIELY. « Frontière linguistique et destin de la Belgique ». In : *Politique étrangère* 47.3 (1982), p. 659.

7. RILLAERTS, « La frontière linguistique, 1878-1963 », p. 16-27.

8. Dave SINARDET. « Territorialité et identités linguistiques en Belgique ». In : *Hermès* n° 51.2 (2008), p. 141-142.

9. JAVEAU, « L'énigme de la frontière linguistique en Belgique », p. 49.

10. RILLAERTS, « La frontière linguistique, 1878-1963 », p. 28-30.

fois de suivre un enseignement entièrement en néerlandais.¹¹ L'unilinguisme territorial est une volonté du mouvement wallon, qui ne souhaite pas un bilinguisme généralisé, mais qui par cette décision abandonne les élites francophones en Flandre.¹²

La frontière linguistique, qui évoluait selon les recensements linguistiques décennaux, est fixée définitivement dans le cadre des lois linguistiques de 1962-1963 (lois Gilson), qui complètent la loi de 1932.¹³ La crainte du « débordement francophone » et de l'extension de la « tache d'huile » autour de Bruxelles est la motivation principale de cette fixation, voulue par le mouvement flamand.¹⁴

Nous pouvons voir la fédéralisation du pays sur base d'une distinction avant tout linguistique, à partir de 1970, comme l'aboutissement de ces différentes lois précédemment mentionnées¹⁵ : tout se sépare selon les lignes linguistiques précédemment tracées, notamment les partis politiques ou les universités (l'ULB et la VUB se séparent en 1970). On aboutit à un état de forte polarisation entre les nouvelles entités fédérées, qui contribuera au caractère toujours évolutif du fédéralisme belge.¹⁶

2.2 Récupération des textes sur CAMille

Nous avons choisi d'utiliser les données issues de trois journaux de tendances politiques différentes :

- La Libre Belgique, de tendance catholique et qui est par ailleurs le seul périodique francophone qui dispose d'un réel lectorat en Flandre¹⁷ ;
- Le Peuple, qui est le quotidien d'information du Parti socialiste belge et suit fidèlement la ligne du parti (il est véritablement un de ses organes et n'en est pas indépendant)¹⁸ ;
- Le Soir, de tendance indépendante et qui revendique sa neutralité tout en étant surtout lu par la moyenne bourgeoisie francophone.¹⁹

En ce qui concerne l'équation de recherche, cette dernière est somme toute assez simple : « question linguistique ». L'utilisation de cette formulation nous a été inspirée d'un article de JASSEM-STANIECKA²⁰ : cette expression fait références aux tensions culturelles, politiques, institutionnelles entre les groupes linguistiques du pays. Nous avons utilisé des guillemets pour capturer l'expression dans son intégralité. Cette façon de procéder a le défaut d'exclure toute imperfection de l'OCR, ce qui n'est pas rare dans le corpus de CAMille. Néanmoins, l'utilisation d'une recherche approximative basée sur la

11. GORIELY, « Frontière linguistique et destin de la Belgique », p. 661, 667 ; RILLAERTS, « La frontière linguistique, 1878-1963 », p. 31-40.

12. SINARDET, « Territorialité et identités linguistiques en Belgique », p. 143 ; WITTE, « La question linguistique en Belgique dans une perspective historique », p. 45.

13. SINARDET, « Territorialité et identités linguistiques en Belgique », p. 144 ; WITTE, « La question linguistique en Belgique dans une perspective historique », p. 45-46 ; pour un compte-rendu exhaustif sur ces lois qui ne rentrent pas dans le cadre de notre corpus, voir RILLAERTS, « La frontière linguistique, 1878-1963 », p. 55-87.

14. JAVEAU, « L'énigme de la frontière linguistique en Belgique », p. 51 ; RILLAERTS, « La frontière linguistique, 1878-1963 », p. 40-54.

15. WITTE, « La question linguistique en Belgique dans une perspective historique », p. 46, 50.

16. GORIELY, « Frontière linguistique et destin de la Belgique », p. 670 ; SINARDET, « Territorialité et identités linguistiques en Belgique », p. 145.

17. René CAMPÉ, Marthe DUMON et Jean-Jacques JESPERS. *Radioscopie de la presse belge*. Verviers : Marabout, 1975, p. 58-59.

18. CAMPÉ, DUMON et JESPERS, *Radioscopie de la presse belge*, p. 98-99.

19. CAMPÉ, DUMON et JESPERS, *Radioscopie de la presse belge*, p. 148-149.

20. Anna JASSEM-STANIECKA. « Community Conflict in Belgium and its Linguistic Reflections ». In : *ALPPI Annual of Language & Politics and Politics of Identity* 6 (2012), p. 3.

distance de Levenshtein est impossible : les termes « question » et « linguistique » sont trop courants et ne pas utiliser l’expression exacte grâce aux guillemets amène trop de bruit dans les résultats renvoyés par la recherche.

Le dernier filtre que nous avons appliqué n’en est pas vraiment un. En effet, nous avons choisi de sélectionner uniquement les résultats entre 1919 et 1950. Le deuxième date est évidente, puisqu’elle marque la fin des données pour *La Libre Belgique* et *Le Peuple*. Pouvoir inclure les développements postérieurs autour de la « question linguistique » aurait été pertinent, mais l’étendue des données nous a empêché de l’envisager. La première a été sélectionnée puisque cette date marque le début de résultats conséquents en réponse à l’équation de recherche, et correspond par ailleurs à la prise d’importance du conflit linguistique dans la vie politique belge (voir ci-dessus). L’absence de résultats antérieurs peut s’expliquer par la mauvaise qualité de l’OCR dans de plus anciens journaux, qui n’a pas pu être compensée dans notre équation de recherche. La période temporelle ainsi retenue ne permet pas d’englober l’ensemble de la « question linguistique », qui est encore très présente dans la vie politique belge dans les années 1950 et 1960, mais elle permet néanmoins d’étudier une période centrale dans ce conflit (avec les années 1920 et 1930 notamment).

3 Exploration préliminaire des données

Puisqu’un des objectifs que nous avons identifiés pour ce travail est de comparer les résultats suivant les journaux, nous serons souvent amenés à considérer à la fois le corpus dans son intégralité et le corpus divisé en journaux. Ceci se reflète dès le début de notre traitement, en créant un fichier `.txt` qui regroupe l’ensemble du corpus, et trois fichiers `.txt` contenant chacun les données d’un journal.

Le corpus global comprend 2035 fichiers, parmi lesquels 700 proviennent de *La Libre Belgique*, 553 du *Peuple* et 782 du *Soir*. En somme, un corpus assez équilibré. Toutes les années de la période 1919-1950 sont présentes dans le corpus, même si la deuxième guerre mondiale a vu l’interruption des publications de *La Libre Belgique* et du *Peuple* (absence des années 1941 à 1943 pour *Le Peuple*, 1941 à 1944 pour *La Libre Belgique*).

La répartition des fichiers par année est similaire dans le corpus global (voir la figure 1) ou dans chaque journal (voir la figure 2). On observe que les années 1919-1925, 1929-1933 et 1935-1940 sont plus productives que les autres. Les deux premières ne sont guère surprenantes, puisqu’elles correspondent aux lois Van Cauwelaert et Jaspar. Nous remarquons que la période de la deuxième guerre mondiale et encore davantage de l’après-guerre n’a pas produit de nombreux articles sur la question linguistique. C’est un peu surprenant, puisqu’un recensement linguistique a lieu en 1947 et que se prépare une loi en 1954, mais la littérature mentionne de fait que le sujet principal de l’après-guerre est la question royale et non la question linguistique.²¹

Le corpus étudié est très vaste : il compte plusieurs millions de mots, même après un tri rudimentaire pour enlever un maximum de bruit. Néanmoins, comme le montre l’impression des hapax, il reste encore de nombreux mots doux tels que « cenlimesiie », « jstiition » ou encore « fifdaeliöh ». L’impression des mots les plus courants nous permet déjà d’avoir une idée du contenu des fichiers : « gouvernement », « ministre », « bruxelles », « pays »... Notre corpus traite, de façon peu surprenante, de politique.

21. RILLAERTS, « La frontière linguistique, 1878-1963 », p. 40.

4 Mots-clés

Nous avons décidé d'extraire les mots-clés de trois sous-ensembles : le corpus complet, le corpus divisé en journaux, ainsi que les fichiers issus des trois périodes qui ressortent du graphique 1. Pour ce faire, nous avons créé des fichiers `.txt` regroupant toutes les données issues de ces années, en suivant la même procédure utilisée pour créer les fichiers de chaque périodique. Grâce à YAKE, nous pouvons extraire les mots-clés de ces fichiers : nous avons instancié un extracteur qui renvoie les 50 premiers mots-clés. Dans ces mots-clés, nous retenons les bigrammes et trigrammes, c'est-à-dire des groupes de deux ou trois mots, pour plus de lisibilité et afin d'obtenir des résultats qui font sens.

Du corpus complet ressortent les mots-clés suivants (bigrammes puis trigrammes) : van cauwelaert, van zeeland, gouvernement belge, Congo belge... ; gouvernement van zeeland, parti ouvrier belge, paul van zeeland, jour libre belgique, mgr van roey, parti catholique belge... Ce sont des mots-clés attendus pour des extraits de presse de nature politique : des noms de politiciens, de partis, de gouvernements... Van Zeeland fut le premier ministre belge de 1935 à 1937, Van Cauwelaert la figure de proue du mouvement flamand qui a donné son nom à la loi linguistique de 1921, Mgr Van Roey l'archevêque de Malines.

De façon générale, nous observons que les mots-clés qui étaient apparus en premier lieu pour le corpus reprenant tous les fichiers du Soir ne figurent pas parmi nos mots-clés : les petites annonces, les logements à vendre ou à louer, etc. Les pages de journal sélectionnées par l'équation de recherche sur CAMille semblent bel et bien provenir de la section politique des périodiques et ne rien comporter d'autre.

4.1 Par journal

Même si les mots-clés issus de chaque journal ne sont pas foncièrement différents de ceux que nous avons évoqué ci-dessus (ils sont toujours de nature politique et évoquent des gouvernements, des politiciens...), certains d'entre eux nous semblent pertinents :

- La Libre Belgique : libre belgique, parti catholique, catholique belge ; libre belgique annonces, jour libre belgique, cardinal van roey, l'union catholique belge, jeunesse catholique belge, parti catholique belge, ministres catholiques flamands...
- Le Peuple : parti socialiste, parti ouvrier ; parti ouvrier belge, ouvrier belge bruxelles, parti socialiste belge, parti socialiste français, fédération syndicale internationale...
- Le Soir : Congo belge, société belge, affaires étrangères, conseil général ; ligue nationale belge, société royale belge, roi george londres...

Remarquons que certains des mots-clés reflètent le position idéologique du journal. Dans La Libre Belgique, les mots-clés liés à « catholique » (parti, union, jeunesse, ministres) sont très fréquents. Dans Le Peuple, c'est au contraire des mots-clés liés à « socialiste » ou « ouvrier », parfois par-delà les frontières de la Belgique. En ce qui concerne Le Soir, nous constatons une absence des mots-clés liés aux idéologies parmi les mots les plus fréquents, et nous trouvons à la place des considérations plus générales (ligue nationale belge, conseil général) ou liées à l'étranger (roi george londres, ainsi que de nombreuses mentions des capitales européennes). Signalons également que La Libre Belgique est le seul journal où le nom du périodique lui-même apparaît très vite dans les mots-clés.

4.2 Focus sur trois périodes

La période 1935-1940 est caractérisée par les mots-clés suivants : van zeeland, défense nationale, parti catholique, van cauwelaert ; gouvernement van zeeland, parti ouvrier belge, bloc catholique belge, cabinet van zeeland... Le nom de Van Zeeland est très

présent, ce qui est normal puisqu'il fut premier ministre de 1935 à 1937. Celui de Van Cauwelaert peut être un peu plus surprenant, puisqu'il est à l'origine de la loi de 1921. En réalité, il est resté une figure de proue du mouvement flamand et devient même en 1939 le chef de file du groupe flamingant au parlement, ce qui expliquerait aisément sa présence dans les mots-clés de la période. Le parti catholique prend le nom de bloc catholique en 1936, d'où la présence du mot-clé correspond.

La période 1919-1925 est caractérisée par les mots-clés suivants : van cauwelaert, parti catholique, gouvernement allemand, parti socialiste, lloyd george ; parti ouvrier belge, parti socialiste belge, parti catholique belge, union catholique belge... Les élections de 1919 voient une nette avancée du parti socialiste et un recul du parti catholique, qui va se restructurer en union catholique en 1921. C'est aussi la période de l'après-guerre, d'où pourraient provenir la raison des références à l'étranger (gouvernement allemand, premier ministre britannique Lloyd George...). La présence de Van Cauwelaert est plus attendue ici, puisque c'est la période de la loi qui porte son nom.

La période 1929-1933 est caractérisée par les mots-clés suivants : van cauwelaert, société nationale, conseil général ; parti ouvrier belge, bruxelles chambre catholiques, ligue nationale belge, ministres catholiques flamands, mgr van roey, cardinal van roey... Van Cauwelaert demeure présent dans cette période, preuve de la longue relation qu'il a entretenue avec la question linguistique belge. Le nom de Van Roey apparaît : il est nommé archevêque de Malines en 1926 et créé cardinal en 1927. Il n'est donc guère étonnant qu'il soit présent dans la presse quelques années après : dans un pays où le parti catholique reste encore au pouvoir, le primat de l'Église belge est une personne importante.

5 Nuages de mots

Pour créer les nuages de mots, nous avons appliqué une fonction de nettoyage sur le fichier reprenant l'ensemble du corpus et sur les trois fichiers avec les données de chaque journal. Cette fonction a été complétée par une liste de *stopwords* comprenant des mots que nous estimons vides de sens (les, plus, cette, fait, faire...), des mots qui ne sont pas pertinents dans le contexte du journal (agence, nord, midi, royale, ville...) et des coquilles et mots sans sémantique discriminante ajoutés après avoir réalisé une première fois les nuages de mots (heures, grand, jour, lieu, temps...).

Les mots qui apparaissent sur le nuage de mots global sont attendus (voir la figure 3) : bruxelles, gouvernement, pays, ministre, belgique, politique, parti, question, conseil... Bref, des mots liés à un contexte politique dont nous sommes devenu bien familier.

Alors que les mots-clés de chaque journal nous donnaient des informations supplémentaires sur ceux-ci, ce ne semble pas être le cas des nuages de mots. En effet, les mêmes mots apparaissent pour le nuage de mots de chaque périodique (voir la figure 4). À titre d'exemples, voici les dix mots les plus fréquents (qui seront les plus grands sur le nuage de mots) :

- La Libre Belgique : gouvernement, ministre, question, pays, belgique, bruxelles, président, conseil, parti, politique ;
- Le Peuple : gouvernement, bruxelles, question, pays, ministre, peuple, parti, politique, belgique, travail ;
- Le Soir : gouvernement, ministre, bruxelles, pays, président, belgique, question, général, paris, prix.

Si ce n'est remarquer que les mots « travail » et « peuple » (ainsi que « socialiste » et « socialistes ») ressortent beaucoup plus sur le nuage de mots du Peuple que sur ceux des deux autres journaux, il n'y a pas grand-chose à commenter.

6 Reconnaissance d'entités nommées

Nous avons effectué une reconnaissance d'entités nommées sur les mêmes sous-ensembles de données que pour les mots-clés : le corpus complet, le corpus journal par journal, une sélection du corpus selon trois périodes temporelles.

Voici les personnes, lieux et organisations qui apparaissent le plus souvent dans le corpus complet :

- Personnes : Henry Bordeaux (5x), président Wilson (5x), Adler (3x), M. von Kardorff (3x), Lénine (2x)...
- Lieux : Allemagne (26x), Bruxelles (16x), Belgique (12x), Paris (9x), Angleterre (8x)...
- Organisations : Académie (3x), Sénat (2x), Chambre (2x)...

Les résultats sont étranges. Parmi les personnes, aucune mention des ministres, rois, politiciens apparaissant dans les mots-clés. On y retrouve, par exemple, un écrivain et académicien français ou un président américain. Nous ne savons pas réellement comment interpréter ou exploiter ces résultats.

6.1 Par journal

Nous avons ensuite appliqué la même procédure sur chaque journal, dont voici les résultats :

- La Libre Belgique : personnes : Henry Bordeaux, président Wilson, Adler... ; lieux : Allemagne, Bruxelles, Belgique... ; organisations : Académie, Sénat, Chambre... ;
- Le Peuple : personnes : M. Braun, Mme Henrickx, M. Vandeperre... ; lieux : Belgique, Bruxelles, Flandre... ; organisations : Chambre, Parlement, Société des Nations... ;
- Le Soir : personnes : Reiss, Foch, M. Wilson, C. Huysmans, M. Mechelynck... ; lieux : Belgique, Allemagne, Bruxelles, Paris... ; organisations : Chambre, Parlement, Société des Nations...

Les lieux et les organisations sont similaires pour chaque journal. Le Peuple semble néanmoins surtout contenir des lieux belges, contrairement aux deux autres qui évoquent des pays et capitales étrangères. Les personnes reconnues sont différentes mais difficile à identifier. Le Soir semble reconnaître Camille Huysmans, un premier ministre belge, mais nous n'avons pas pu aller plus loin dans la reconnaissance des personnes.

6.2 Focus sur trois périodes

Enfin, nous avons effectué la reconnaissance sur les trois périodes sélectionnées. Voici les résultats pour chacune d'entre elles :

- 1919-1925 : personnes : Henry Bordeaux, président Wilson, Adler... ; lieux : Allemagne (26x!), Bruxelles, Belgique... ; organisations : Académie, Sénat, Chambre... ;
- 1929-1933 : personnes : Jeanne, M. Brifaut, pape... ; lieux : Belgique, Allemagne, Berlin, Bruxelles... ;
- 1935-1940 : personnes : M. Meyer, Hitler, Krahmer... ; lieux : la Sarre, Allemagne, Bruxelles, Belgique... ; organisations : Conseil, Reich...

Comme pour les journaux, certaines personnes évoquées sont difficiles à identifier : citons Jeanne ou M. Meyer. Nous en reconnaissons néanmoins quelques-unes : Henry Bordeaux, précédemment cité, devient académicien en 1919 et Wilson est lui aussi président des États-Unis pendant une partie de la période considérée. Tous deux apparaissent dans le groupe des années 1919-1925. De même, Hitler est évoqué pour la période 1935-1940,

qui correspond à ses premières années de pouvoir en Allemagne. Au niveau des lieux, nous remarquons que l'Allemagne est beaucoup évoquée dans la première période, celle de l'après-guerre (plus du double du deuxième lieu le plus mentionné, Bruxelles), et le redevient pour la troisième période, aux côtés de la Sarre qui que l'Allemagne récupère en 1935. Enfin, Reich apparaît dans les organisations pour la troisième période, autre preuve de la place grandissante de l'Allemagne pendant cette période, en dépit du fait que seules les pages traitant de la question linguistique belge ont été sélectionnées et téléchargées de CAMille.

7 Clustering

Pour pouvoir effectuer un clustering efficace, les textes sont tokenisés et la ponctuation est enlevée. Ils sont ensuite transformés en vecteurs grâce au modèle TF-IDF. Nous avons, dans un premier temps, effectué un clustering sur le corpus complet. Au vu des résultats obtenus, et même si nous ne pensions pas le faire, nous avons appliqué la même procédure aux textes de chacun des trois périodiques de notre corpus. Dans les deux cas, nous avons exploré les clusters en extrayant leurs mots-clés et en observant les mots les plus fréquents (sans pour autant créer de nuage de mots).

7.1 Sur le corpus complet

L'application de l'*elbow method* indique très nettement que le paramétrage idéal consiste en deux clusters (voir la figure 5). Une analyse en composantes principales permet de visualiser qu'en effet, ce paramétrage semble efficace (voir la figure 6).

Les résultats du clustering étant un dictionnaire, nous en avons exploré le contenu grâce à une boucle basée sur les *keys* du dictionnaire. Le premier cluster comporte 1255 documents, le deuxième 780. La répartition du nombre de documents par année semble similaire à celle observée pour le corpus dans son entièreté (voir la figure 7). L'année n'a donc pas été un facteur déterminant dans le cadre de cette opération de clustering. Les mots-clés extraits ne nous renseignent pas davantage. Les mots-clés trouvés dans le cluster 0 sont : société nationale, défense nationale, roi albert, ligue nationale belge, gouvernement van zeeland, société royale belge, roi george londres... Ceux dans le cluster 1 sont : van cauwelaert, parti catholique, van zeeland, gouvernement belgique, parti socialiste, peuple flamand ; parti ouvrier belge, gouvernement van zeeland, anvers bruxelles gand, libre belgique annonces, bloc catholique belge... Nous n'avons pas trouvé de séparation évidente au niveau des thématiques que ces mots-clés évoquent.

Néanmoins, puisque nous avons étudié les mots-clés de chaque journal, nous avons remarqué des similitudes avec des observations précédentes et avons décidé de visualiser les journaux qui se retrouvent dans chaque cluster (voir la figure 8). La conclusion à tirer de ces graphiques est claire : le cluster 0 regroupe l'écrasante majorité des fichiers du Soir, tandis que le cluster 1 comprend les fichiers du Peuple et de La Libre Belgique (ainsi que deux fichiers isolés du Soir). Ainsi, le clustering a divisé le corpus selon les journaux. Armé de cette connaissance, nous avons tenté de reproduire le clustering avec trois clusters, en supposant que les trois clusters allaient correspondre aux trois journaux. L'hypothèse s'est révélée erronée (comme pouvait le suggérer l'Elbow Method, qui n'indique pas un clustering efficace avec trois clusters). Faudrait-il donc en conclure que Le Soir possède un je-ne-sais-quoi qui le sépare des deux autres journaux : sa fameuse neutralité souvent revendiquée ? De fait, on retrouve beaucoup moins de mots-clés liés aux partis et idéologies dans le cluster du Soir (cluster 0) que dans l'autre cluster.

7.2 Sur le corpus de chaque journal

Puisque le clustering du corpus global s’est réalisé selon les journaux, il nous a semblé intéressant d’observer rapidement les résultats d’un clustering sur les fichiers de chaque journal séparément. Pour ce faire, nous avons créé une liste avec les trois matrices de vecteurs issues de l’application de TF-IDF sur les trois fichiers des journaux. La méthode employée est donc la même que pour le corpus complet, avec l’application d’une boucle pour étudier l’ensemble des trois matrices de façon simultanée.

Nous ne rentrerons pas dans les détails de chaque périodique, car nous pensons qu’une observation globale suffit. En annexe se trouvent les graphiques de l’Elbow Method (figure 9) ainsi que les visualisations des clusters et la répartition en année pour chacun des trois journaux (figures 10, 11 et 12). Les graphiques appliquant l’Elbow Method nous montrent que le meilleur nombre de clusters n’est pas évident à choisir : nous avons choisi, sur base de ces graphiques et des visualisations par analyse en composantes principales (et donc de façon subjective), de préférer deux ou trois clusters.

Il est difficile de tirer des observations de ce clustering : les graphiques de répartition des documents par année se ressemblent de cluster en cluster, de même que les mots-clés et mots les plus fréquents. Prenons l’exemple du Soir, divisé en trois clusters (ce qui semble faire sens au vu de l’analyse en composantes principales). Le nombre de documents par année est similaire au corpus global, avec des pics sur trois périodes distinctes. Le cluster 2 est quelque peu différent, puisqu’il ne comprend pas toutes les années (contrairement aux deux autres). Les mots-clés de chaque cluster sont : libre belgique, van cauwelaert, parti catholique, gouvernement belge (cluster 0) ; libre belgique, van cauwelaert, gouvernement belge (cluster 1) ; libre belgique, dans tous, van cauwelaert (cluster 2). De même, les mots les plus fréquents : gouvernement, ministre, question (cluster 0) ; gouvernement, ministre, question (cluster 1) ; gouvernement, ministre, question (cluster 2). En d’autres termes, nous pouvons supposer que le contenu thématique de chacun des trois clusters se ressemble entre eux à s’y méprendre, et que par conséquent le clustering est inefficace pour nous donner des informations complémentaires sur le corpus divisé en journaux.

Enfin, soulignons une étrangeté au niveau du traitement du Peuple. Pour une raison qui nous échappe, l’exploration des clusters révèle que seuls les fichiers de 1919 à 1936 sont considérés. Or, une impression de la liste `tfidf_vectors_LP` nous confirme qu’elle comprend bel et bien les fichiers jusqu’à la fin de notre corpus, c’est-à-dire 1950. Nous n’y voyons pas d’explication, surtout que le code appliqué sur les fichiers du Peuple est exactement similaire à celui utilisé pour La Libre Belgique ou Le Soir.

8 Plongements lexicaux

Nous avons décidé d’utiliser les plongements lexicaux (*word embeddings*) pour souligner des différences entre les trois corpus de périodique. Pour ce faire, nous avons défini une fonction `sentence_tokenizer` qui permet de scinder les fichiers en phrases. Après un très long traitement (plus de quatre heures), nous obtenons trois nouveaux fichiers que nous pouvons utiliser pour créer les modèles de plongements lexicaux.

8.1 Création des modèles word2vec

Comme nous l’avons observé lors du TP3 en nous appuyant sur la littérature scientifique, les paramètres `min_count` et `window` permettent de modifier le type d’associations réalisées par word2vec : un petit nombre favorisera les associations syntaxiques, alors

qu'un plus grand mettra en évidence des relations sémantiques.²²

Puisque nous avons déjà trois corpus à entraîner et observer (un par journal), nous avons décidé de créer deux modèles : un premier avec les paramètres par défaut (`min_count` = 5 et `window` = 5) et un deuxième avec `min_count` = 20 et `window` = 20. Dans les observations, le premier modèle sera appelé « modèle syntaxique » et le deuxième « modèle sémantique ». Les deux modèles sont alors insérés dans des listes, qu'il est facile de parcourir à l'aide d'une boucle.

L'entraînement a été étonnamment rapide, ne durant que quelques minutes. Nous avons eu l'habitude d'entraîner des modèles sur l'ensemble du corpus du Soir, qui est bien plus volumineux que les corpus sur lesquels nous travaillons ici, même s'ils sont au nombre de trois.

8.2 Observations

La fonction `similarity` permet de calculer la similarité entre deux termes. Nous avons sélectionné un ensemble de couples de mots qui, d'après nos lectures dans la littérature scientifique, devraient être plus ou moins proches l'un de l'autre : « question » et « linguistique », « bilinguisme » et « neerlandais », « bilinguisme » et « francais », « langue » et « enseignement », « université » et « neerlandais », « universite » et « flamand », « catholique » et « majorite », « socialiste » et « majorite », « liberale » et « majorite », « socialiste » et « catholique », « catholique » et « liberal », « liberal » et « socialiste ». Remarquons que les mots ont été tokénisés et doivent par conséquent être écrits sans flexion morphologique (pluriel, féminin...) ou sans signe diacritique (accents, cédille...). Nous emploierons cette graphie dans le cadre de nos observations. Voici quelques commentaires sur les résultats obtenus :

- Les mots de l'expression « question linguistique » sont moins similaires que nous pouvions le supposer, avec une similarité de 0.6 environ (sauf Le Peuple où l'association est plus forte). Ceci peut s'expliquer par le fait que « question » est polysémique ;
- « bilinguisme » est plus similaire à « neerlandais » qu'à « francais », peu importe les journaux et les modèles considérés ;
- « langue » et « enseignement » sont fortement liés dans La Libre Belgique, mais c'est moins le cas dans les autres journaux ;
- Alors que « universite » et « neerlandais » ont un taux de similarité bas, « universite » et « flamand » sont très proches l'un de l'autre, surtout dans Le Peuple ;
- L'association de « majorite » aux idéologies obtient des résultats différents selon les journaux : ce mot est fort associé à « catholique » et « liberal » dans Le Peuple, à « socialiste » dans La Libre Belgique. Cette association est intéressante, puisqu'elle concerne des idéologies qui ne sont pas représentées par le journal en question. Le taux de similarité augmente dans le modèle sémantique, ce qui semble confirmer notre première observation ;
- Les termes « catholique », « liberal » et « socialiste » sont tous très similaires les uns aux autres.

La fonction `most_similar`, quant à elle, ne prend en entrée qu'un seul mot et fournit les mots les plus proches (dans un ordre descendant). Nous avons dressé une liste de quelques mots, qui se recoupe par endroit avec les mots considérés pour la fonction précé-

22. Daniel JURAFSKY et James H. MARTIN. *Speech and Language Processing*. 2021, p. 125 ; Omer LEVY et Yoav GOLDBERG. « Dependency-based word embeddings ». In : *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*. 2014, p. 302-308.

dente : « catholique », « socialiste », « liberale », « liberal », « neerlandais », « francais », « flamand », « wallon », « bruxelles », « langue », « question ». Nous avons intégré à la fois « liberal » et « liberale » car nous avons constaté que, sûrement à cause d'une mauvaise tokénisation, les deux formes existent dans notre corpus. Voici les observations que nous avons pu tirer des résultats obtenus :

- « catholique » est proche de mots comme : liberale, socialiste, communiste, du parti... Bref, d'autres idéologies politiques ;
- « socialiste » est proche de mots comme : communiste, liberal, du parti... La Libre Belgique l'associe d'abord à « communiste », tandis que Le Peuple à « catholique » ou « national ». Peut-on y voir une forme de rejet de la part du premier périodique, de tendance catholique, et une forme d'intégration de l'idéologie dans le champ politique belge par le deuxième périodique, fermement socialiste ? Cette distinction est gommée par le modèle sémantique, ce qui laisse supposer qu'il s'agit d'une différence superficielle ;
- « liberal » ou « liberale » sont rapprochés de mots comme : sénateur, rapporteur, député (La Libre Belgique) ; catholique, communiste (Le Peuple). Le Soir l'associe plutôt à des noms ou des fonctions (delattre, ministre transports, houtart...). Le modèle sémantique fait disparaître les différences de résultats entre « liberal » et « liberale » : le modèle syntaxique donne des adjectifs synonymes qui s'éloignent du champ lexical de la politique, tandis que le modèle sémantique propose des synonymes plus diversifiés ;
- « neerlandais » est proche de mots comme : service technique, paysan, hollandais (La Libre Belgique) ; universel, groupe flamand, tact (Le Peuple) ; parti économique, parti catholique (Le Peuple) ;
- « francais » est proche de mots comme : anglais, flamand (La Libre Belgique) ; anglais, flamand, allemand, wallon (Le Peuple) ; anglais, flamand, allemand, hollandais (Le Soir). Ce mot désigne à la fois la langue française mais aussi le peuple français, expliquant ces résultats hétéroclites ;
- « flamand » est proche de mots comme : wallon, parti catholique, peuple, pays flamand (La Libre Belgique) ; wallon, parti, socialisme, francais, catholique (Le Peuple) ; allemand, francais, parti, danger, bloc, wallon, wallons (Le Soir). La Libre Belgique utilise des termes qui font référence à la population et pas seulement à la politique (peuple, pays), à mettre en lien avec sa caractéristique d'être le seul périodique avec un lectorat en Flandre ? Le mot « wallon » est également très présent, sauf dans Le Soir où seul le modèle sémantique l'extrait. Le Soir évoque également des mots comme « danger » ou « allemand », peut-être avec une vision plus négative de la Flandre ? Cette interprétation nous semble néanmoins légère, ne se basant que sur quelques maigres similitudes ;
- « wallon » est proche de mots comme : flamingant, paysan, flamand, citoyen, chrétien (La Libre Belgique) ; américain, code, commissariat général, flamand, député, conservateur (Le Peuple) ; puissant, spirituel, indépendant, paysan, flamand, bourgeois, séparatisme (Le Soir). Ce mot semble avoir des associations plus larges que « flamand », avec notamment des considérations individuelles (avec les qualificatifs qu'on retrouve dans Le Soir). Il est bien sûr également associé à « flamand » ;
- « bruxelles » est proche d'autres villes belges, comme liege, mons, louvain, namur... Il n'est toutefois pas rapproché d'autres capitales européennes ;
- « langue » est proche de mots comme : langue française, culture, nation, communautaire, langue neerlandaise, enseignement (La Libre Belgique) ; mesure, force, démocratie, sécurité, liberté (Le Peuple) ; science, nation, pensée, communautaire,

langue flamande, langue française (Le Soir). Ce mot renvoie à des concepts abstraits (culture, pensée, force...) et, surtout dans le cas du modèle sémantique, aux langues française et flamande (pas néerlandaise!) à proprement parler ;

- « question » est proche de mots comme : réforme, formule, solution, décision... Il n'est pas rapproché de « linguistique », peu importe le journal ou le modèle.

La même fonction `most_similar` permet d'effectuer des recherches plus complexes dans l'espace vectoriel, par exemple en cherchant un mot proche de deux autres mots mais en même temps éloigné d'un autre. Nous avons testé avec quelques mots, sans grand succès. Prenons l'exemple des mots « gouvernement », « catholique » et « socialiste » et cherchons un mot proche des deux premiers mais éloigné du troisième. Les mots les plus proches, selon le modèle syntaxique, sont : parlement, problème, voter, principe, pacte, ministère... Dans le cas du Soir, des réalités étrangères apparaissent : gouvernement français, gouvernement britannique. Pour le modèle sémantique, on obtient des mots comme : devoir, conflit, poursuivre, conclure, voter, maintien, parti libéral, parti catholique... Même si les mots obtenus sont différents selon les journaux et les modèles, nous avons du mal à percevoir en quoi les observations que nous avons pu réaliser pourraient être utiles comme critères de comparaison entre les journaux. En tout cas, nous ne percevons pas de régularités ou de motifs dans les mots obtenus.

Résumons brièvement ce que les plongements lexicaux nous ont appris sur notre corpus. Nous avons pu constater que certains journaux avaient tendance à privilégier un certain vocabulaire. Un exemple parlant à nos yeux est le mot « flamand », que les trois journaux associent différemment (vision incluant le peuple pour La Libre Belgique et perspective uniquement politique pour Le Peuple). Ainsi, cet outil nous permet de découvrir certaines associations que nous ne percevrions pas à l'aide d'une lecture traditionnelle des données. Néanmoins, cette « lecture distante » est limitée. D'une part, les informations qu'elle peut nous donner sur le corpus sont légères et complexes à interpréter. D'autre part, les mots étudiés sont complètement subjectifs, puisque nous devons les sélectionner manuellement. Il y a là un grand risque d'aborder le corpus de façon biaisée.

9 Conclusion

Tout au long de ce travail, nous avons pu constater que les techniques de traitement automatique de corpus nous permettent d'obtenir des informations sur des données qui seraient trop volumineuses pour un traitement manuel efficace en un temps raisonnable (pour rappel, le corpus étudié compte plusieurs millions de mots). Nous avons pu identifier des périodes de discussions intensives autour de la question linguistique, en particulier peu avant le passage de lois mentionnées dans la littérature scientifique (en 1921 et en 1932). Il ne fait aucun doute que, si notre corpus allait jusqu'en 1970, nous verrions apparaître un nouveau pic d'intensité peu avant les lois de 1962-1963.

L'extraction de mots-clés et la création de nuages de mots basés sur la fréquence du vocabulaire ont rendu évident la thématique des données que nous traitons : il s'agit d'articles de nature politique, en particulier sur la politique belge. Nous avons vu apparaître des mots tels que « parlement », « gouvernement », « parti catholique », « parti socialiste », « parti libéral »... Les mots-clés, en particulier, sont efficaces puisqu'ils ont également mis en évidence, en plus des partis ou institutions politiques, certains acteurs de ces conflits politiques : le premier ministre Van Zeeland, le député Van Cauwelaert, le cardinal Van Roey...

En revanche, la reconnaissance d'entités nommées, alors que son objectif est justement d'identifier les personnes dans un corpus, n'a pas été réellement efficace pour repérer

les acteurs principaux de notre corpus et a produit des personnes non identifiées ou d'intérêt mineur. La reconnaissance de lieux ou d'organisations ont elles aussi été d'une utilité limitée, malgré certaines observations intéressantes (des mentions plus importantes de l'Allemagne après la première guerre mondiale ou peu avant la deuxième, aux côtés de Reich et d'Hitler). Ces observations sont toutefois restreintes au contexte historique global et non à notre sujet de recherche. Bref, la reconnaissance d'entités nommées produit soit des résultats trop généraux, comme « Allemagne », « Reich » ou « Société des Nations », ou des résultats tellement précis que non identifiées, comme « Foch » ou « Denis ».

Le clustering a été efficace pour séparer un journal, *Le Soir*, des deux autres. Néanmoins, son efficacité s'est révélée peu satisfaisante puisque, au-delà de cette séparation, il a été impossible de diviser le contenu de chaque journal en clusters pertinents. De plus, le clustering n'a pas permis de créer un cluster par journal, même en changeant les paramètres du clustering sur le corpus complet. Faut-il en conclure que les sujets abordés dans chacun des fichiers sont trop proches les uns des autres et qu'il est impossible de les distinguer ? Il est également pertinent de remarquer que la date des articles n'a aucun impact sur le clustering, peu importe les configurations que nous avons expérimentées.

Les différences entre les trois journaux, soulignées en partie par le clustering, étaient déjà évidentes grâce aux mots-clés. En effet, les mots-clés soulignent les affiliations idéologiques des journaux : *La Libre Belgique*, de tendance catholique, mentionne plus souvent le parti catholique tandis que *Le Peuple*, de tendance socialiste, évoque les ouvriers, le parti socialiste ou encore le parti communiste. L'étude de ces différences a pu être approfondie grâce aux plongements lexicaux, qui a révélé de nombreuses associations différentes entre concepts selon les journaux. Ces résultats sont néanmoins difficiles à synthétiser et à intégrer dans une étude du sujet historique. Nous avons pu remarquer, par exemple, que le mot « majorité » avait tendance à être associé à des idéologies opposées à celles du périodique : *La Libre Belgique* l'associe à « libéral » ou « socialiste », *Le Peuple* à « catholique ». Peut-être qu'un ou une spécialiste de la question pourrait identifier des mots plus pertinents que ceux que le néophyte que nous sommes a sélectionnés et ainsi obtenir des résultats plus pertinents dans le cadre d'une recherche historique ?

L'étude de la perception par la presse belge de la question linguistique bénéficierait sans nul doute d'une extension des données jusqu'en 1970, avec le début de la fédéralisation et d'une nouvelle génération de conflits entre communautés linguistiques. Elle bénéficierait également d'une amélioration de l'OCR, pour deux raisons. D'une part, la qualité des résultats obtenus grâce aux techniques de traitement automatique dépend directement de la qualité des fichiers du corpus. D'autre part, la sélection même des fichiers sur CAMille dépend de l'OCR puisque, pour identifier ceux qui traitent exclusivement de la question linguistique, nous sommes obligés de recourir à une expression exacte qui, par définition, exclut toute erreur d'OCR des résultats obtenus. Ainsi, sur les deux milliers et quelques brouilles de pages de journal qui constituent notre corpus, il est tout à fait envisageable que le même nombre de pages ou plus encore aient été omises lors de la recherche initiale sur la plateforme CAMille. Nous avons pu toutefois démontrer qu'il est possible d'extraire des informations pertinentes de ce corpus imparfait et incomplet, ce qui est tout autant une preuve de la puissance du traitement automatique que de la nécessité de compléter cette approche distante des données avec une expertise humaine du sujet étudié et des documents analysés.

Bibliographie

- [1] Étienne ARCQ, Vincent de COOREBYTER et Cédric ISTASSE. « Fédéralisme et confédéralisme ». In : *Dossiers du CRISP* 79.1 (2012), p. 11-125.
- [2] René CAMPÉ, Marthe DUMON et Jean-Jacques JESPERS. *Radioscopie de la presse belge*. Verviers : Marabout, 1975.
- [3] Georges GORIELY. « Frontière linguistique et destin de la Belgique ». In : *Politique étrangère* 47.3 (1982), p. 657-673.
- [4] Anna JASSEM-STANIECKA. « Community Conflict in Belgium and its Linguistic Reflections ». In : *ALPPI Annual of Language & Politics and Politics of Identity* 6 (2012), p. 23-44.
- [5] Claude JAVEAU. « L'énigme de la frontière linguistique en Belgique ». In : *Revue des Sciences Sociales* 48 (2012), p. 48-53.
- [6] Daniel JURAFSKY et James H. MARTIN. *Speech and Language Processing*. 2021.
- [7] Omer LEVY et Yoav GOLDBERG. « Dependency-based word embeddings ». In : *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*. 2014, p. 302-308.
- [8] Stéphane RILLAERTS. « La frontière linguistique, 1878-1963 ». In : *Courrier hebdomadaire du CRISP* 2069-2070.24 (2010), p. 6-111.
- [9] Jean-Claude SCHOLSEM. « Les métamorphoses du fédéralisme ». In : *Bulletin de la Classe des lettres et sciences morales et politiques* 14.1-6 (2003), p. 15-30.
- [10] Dave SINARDET. « Territorialité et identités linguistiques en Belgique ». In : *Hermès* n° 51.2 (2008), p. 141-147.
- [11] Wilfried SWENDEN et Maarten Theo JANS. « 'Will it stay or will it go?' Federalism and the sustainability of Belgium ». In : *West European Politics* 29.5 (2006), p. 877-894.
- [12] Els WITTE. « La question linguistique en Belgique dans une perspective historique ». In : *Pouvoirs* 136.1 (2011), p. 37-50.

10 Annexes

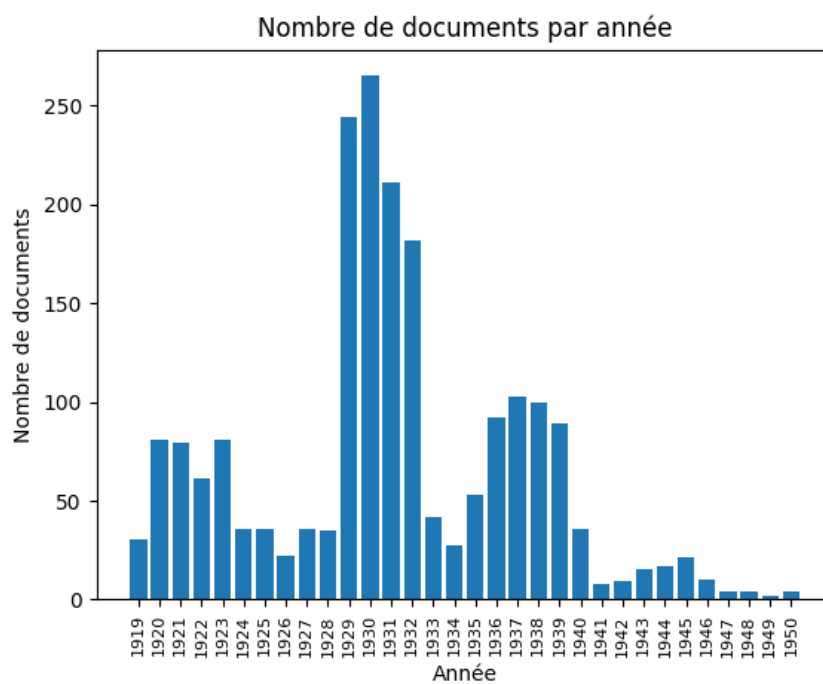


FIGURE 1 – Nombre de documents par année pour le corpus complet

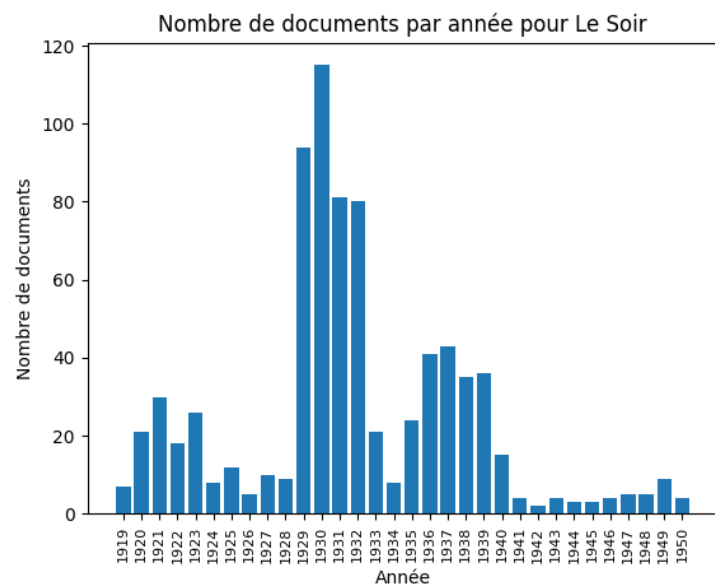
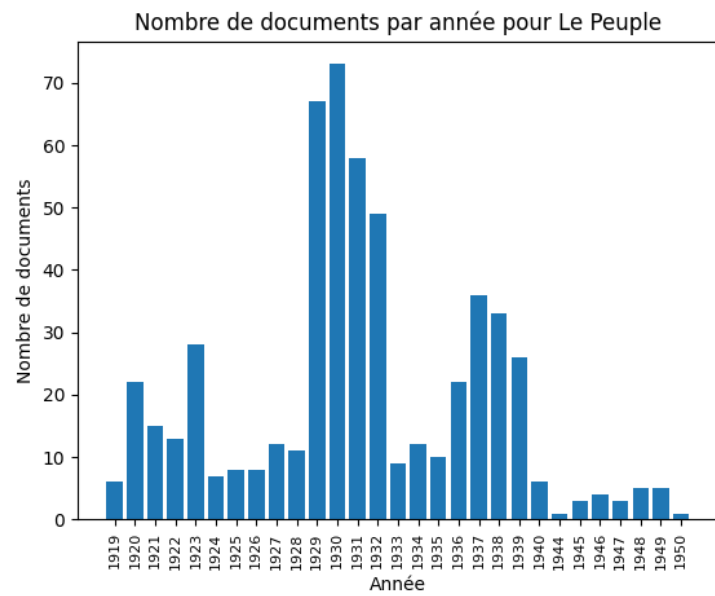
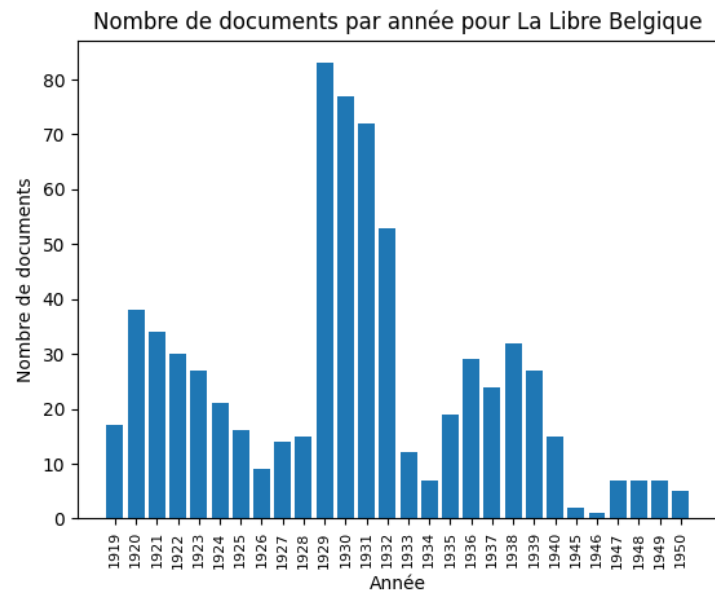


FIGURE 2 – Nombre de documents par année pour chaque journal

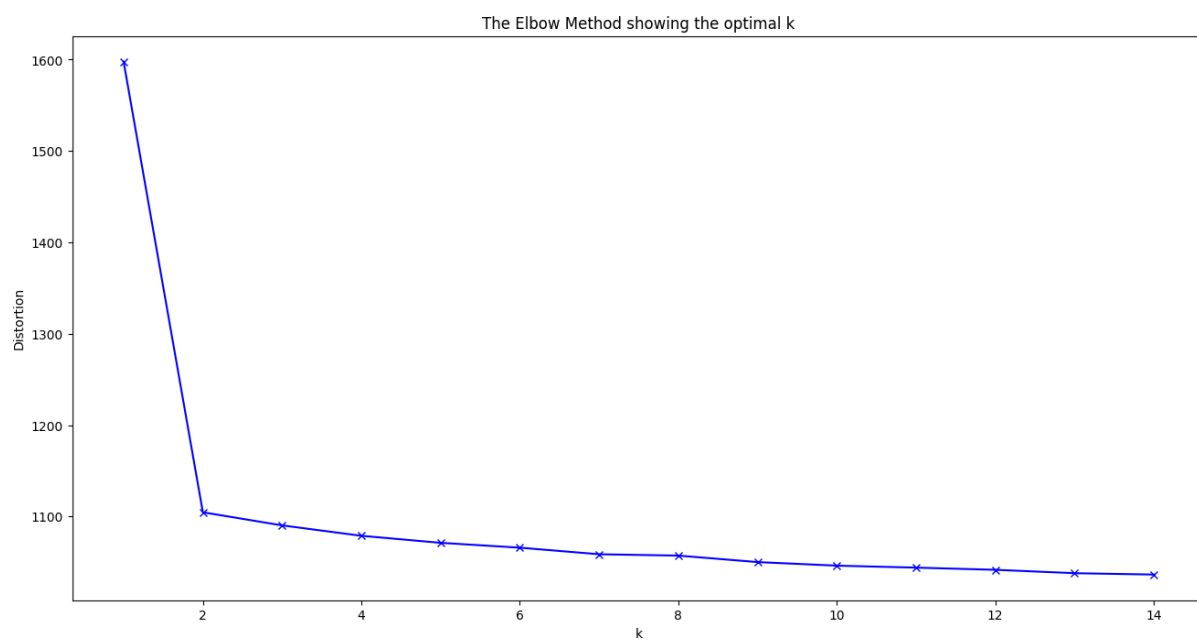


FIGURE 5 – Elbow Method pour le corpus entier

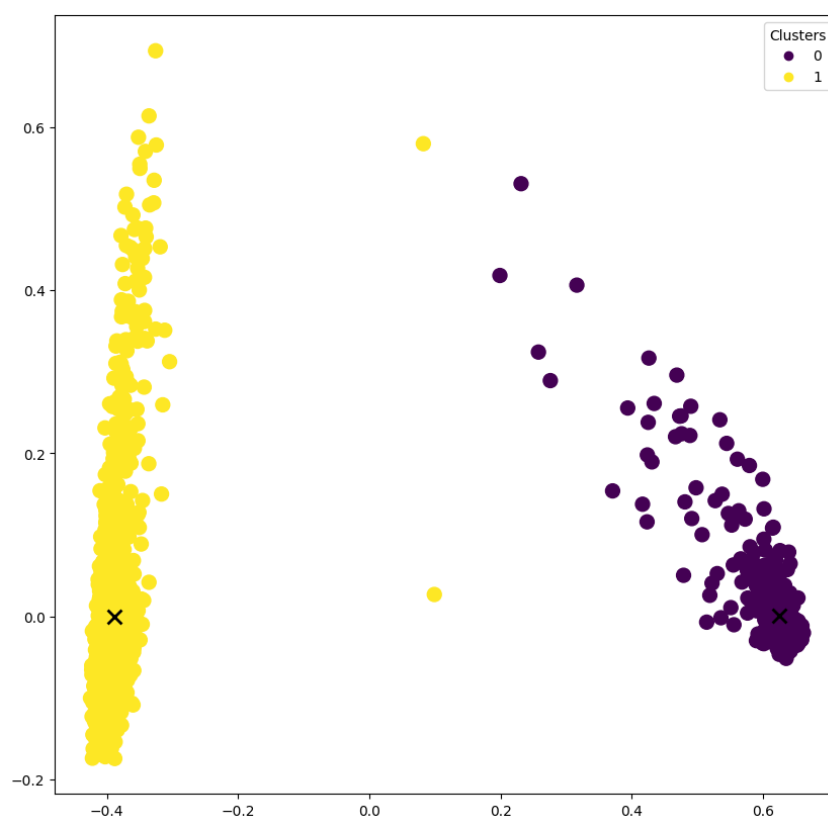
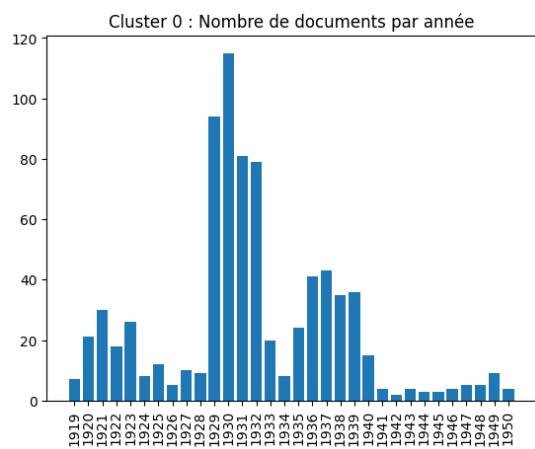
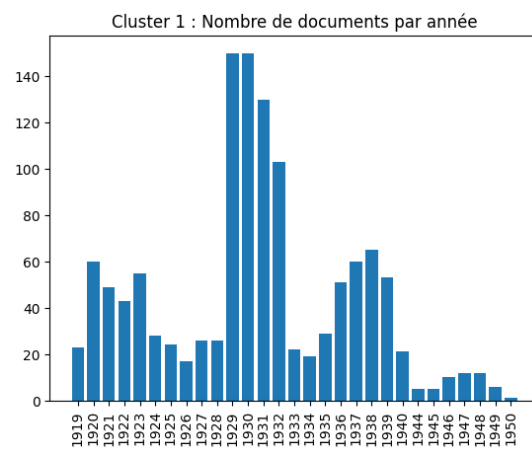


FIGURE 6 – Visualisation du clustering sur le corpus entier (2 clusters)

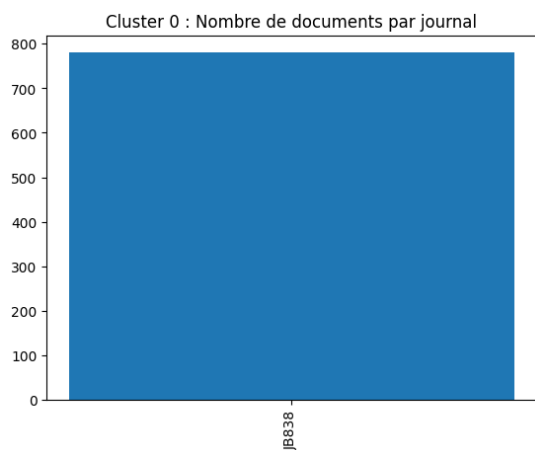


(a) Cluster 0

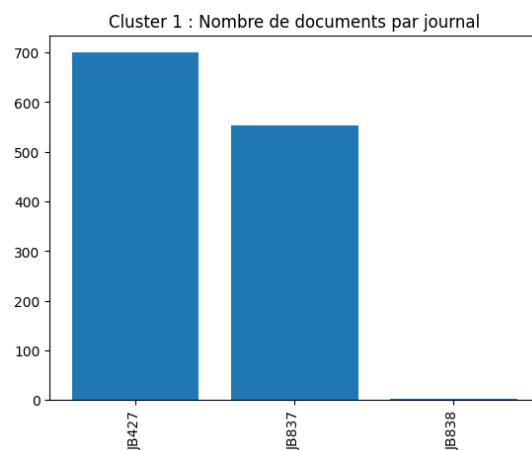


(b) Cluster 1

FIGURE 7 – Nombre de documents par année dans chaque cluster

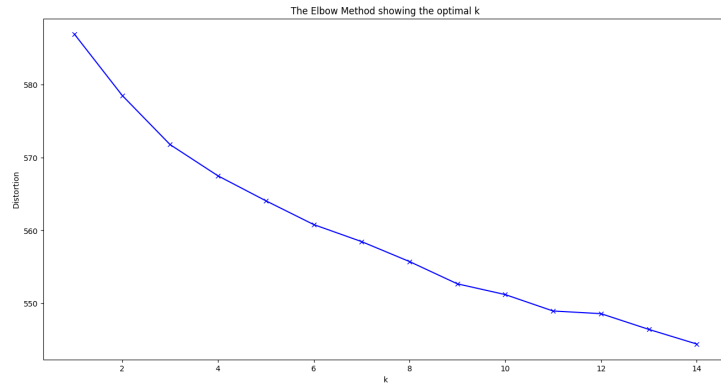


(a) Cluster 0

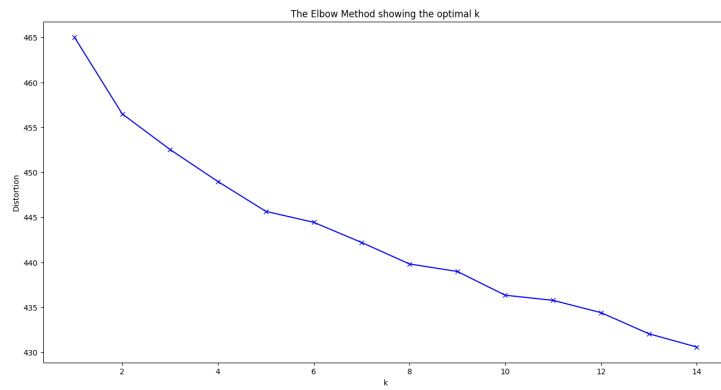


(b) Cluster 1

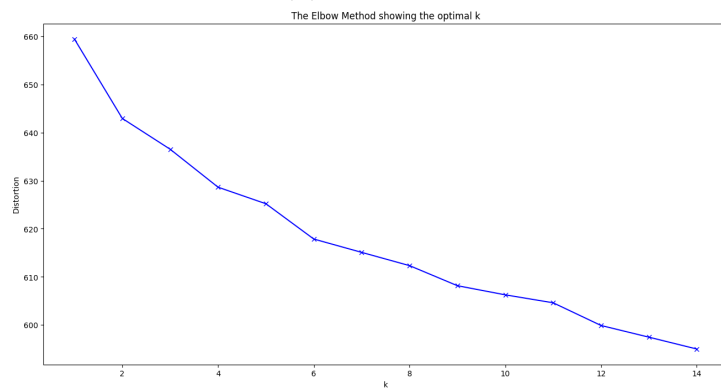
FIGURE 8 – Nombre de documents par journal dans chaque cluster



(a) La Libre Belgique

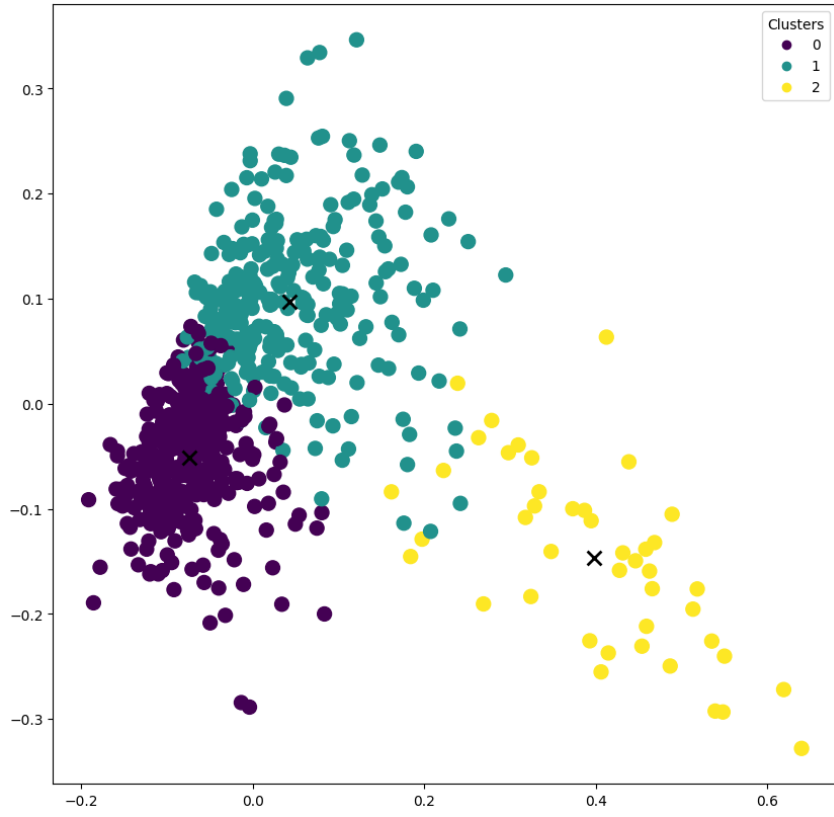


(b) Le Peuple

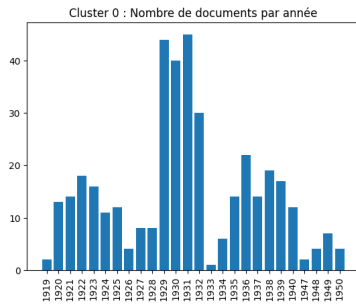


(c) Le Soir

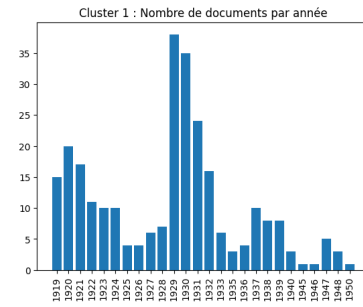
FIGURE 9 – Elbow Method pour chaque journal



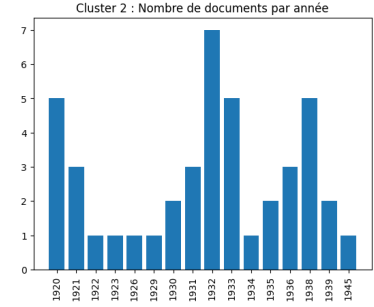
(a) Visualisation du clustering (3 clusters)



(b) Nombre de documents par année – Cluster 0

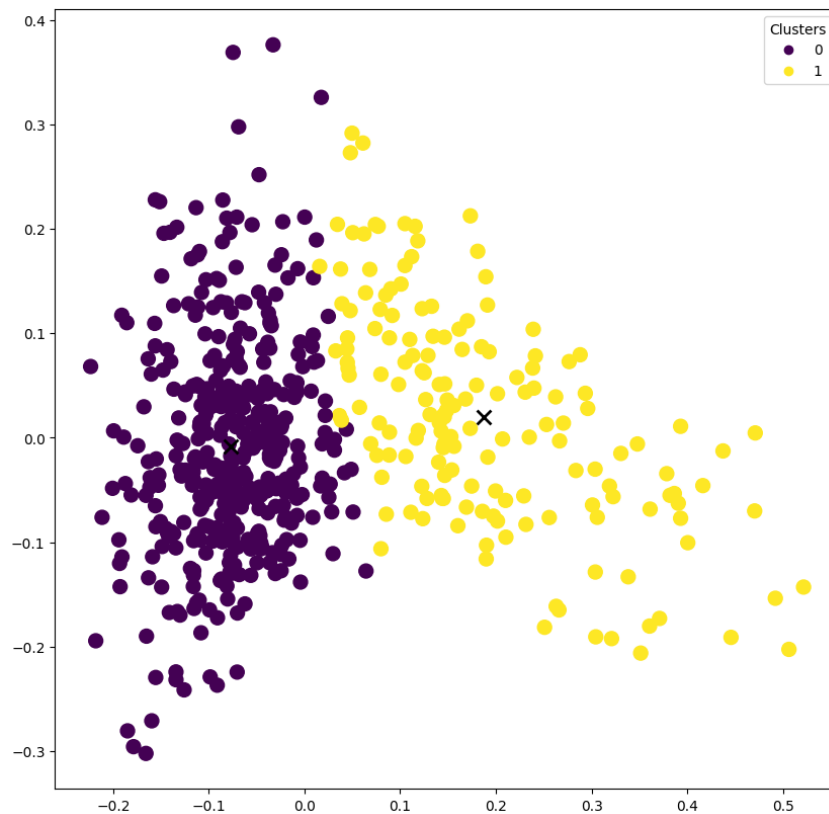


(c) Nombre de documents par année – Cluster 1

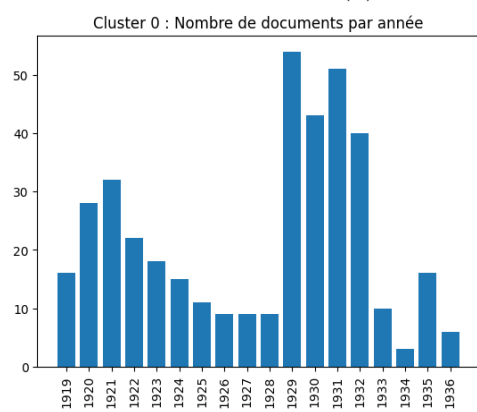


(d) Nombre de documents par année – Cluster 2

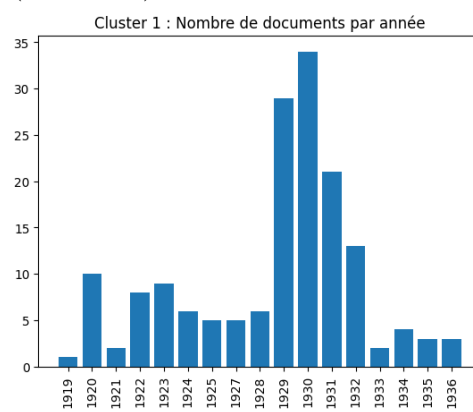
FIGURE 10 – Clustering pour La Libre Belgique



(a) Visualisation du clustering (2 clusters)

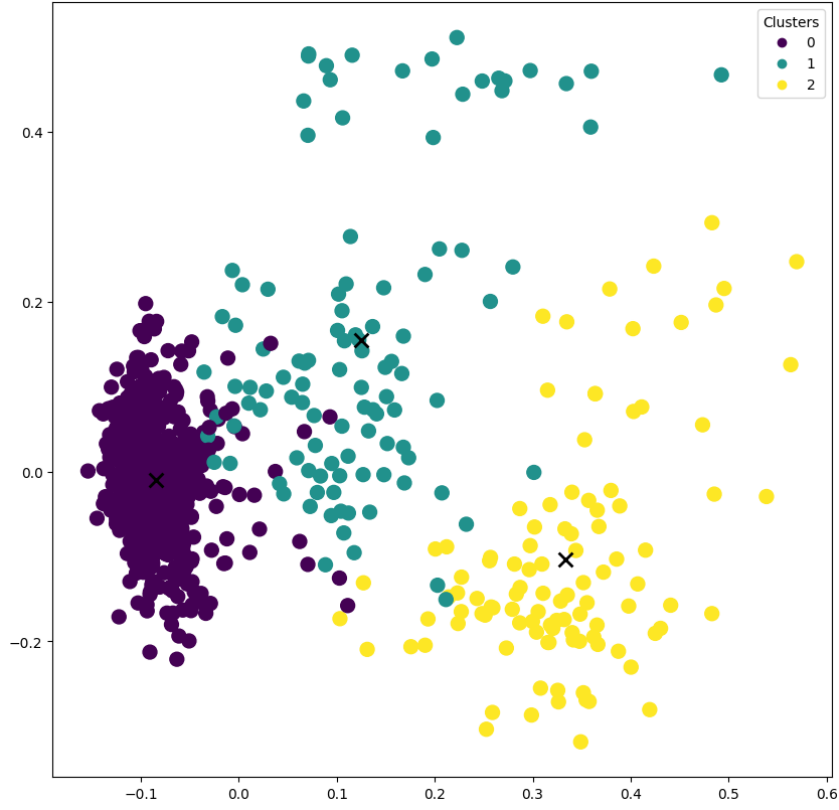


(b) Nombre de documents par année – Cluster 0

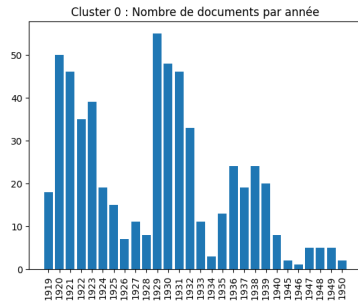


(c) Nombre de documents par année – Cluster 1

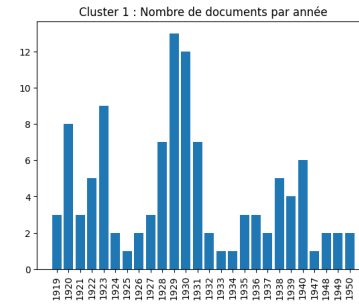
FIGURE 11 – Clustering pour Le Peuple



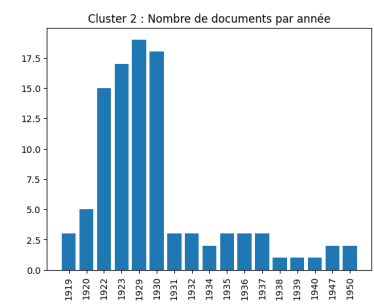
(a) Visualisation du clustering (3 clusters)



(b) Nombre de documents par année – Cluster 0



(c) Nombre de documents par année – Cluster 1



(d) Nombre de documents par année – Cluster 2

FIGURE 12 – Clustering pour Le Soir