Analyser un corpus hétérogène : le cas des Archives numériques de la Révolution française[1]

Pascal Marchand[1], Pierre Ratinaud[2]

[1]Université de Toulouse – pascal.marchand@iut-tlse3.fr

[2] Université de Toulouse – ratinaud@univ-tlse2.fr

Abstract

We analyze a corpus of debates in the National Assembly at the beginning of the French Revolution, from 1789 to 1794, digitized by Stanford University. Initial analyses of this corpus of about 43.4 million words show that digitization is imperfect. We therefore seek to test the heterogeneity of this corpus to see if it is relevant to analyze it. To do so, we build a test sub-corpus that increases the error weight. The results of the hierarchical top-down classification (Reinert, 1983) allow us to analyse the original corpus. Despite the large number of errors caused by digitization, the size of the corpus, the possibility of analyzing a large number of forms and the results of the test sub-corpus allow us to establish the main themes and their chronological distribution. Nevertheless, proposals are made to increase the homogeneity of the corpus as well as of the metadata.

Keywords: archive digitization, heterogeneous corpus, lexical classification, French Revolution.

Résumé

Nous analysons un corpus de débats à l'Assemblée nationale au début de la Révolution française, de 1789 à 1794, numérisé par l'Université de Stanford. Les premières analyses de ce corpus de 43,4 millions de mots environ, montrent que la numérisation est imparfaite. Nous cherchons donc à éprouver l'hétérogénéité de ce corpus pour voir s'il est pertinent de l'analyser. Pour ce faire, nous construisons un sous-corpus de test qui accroît le poids des fautes. Les résultats de la classification hiérarchique descendante (Reinert, 1983) nous autorisent à analyser le corpus d'origine. Malgré le grand nombre de fautes provoquées par la numérisation, la taille du corpus, la possibilité d'analyser un grand nombre de formes et les résultats du sous-corpus de test nous permettent d'en établir les principales thématiques et leur distribution chronologique. Néanmoins, des propositions sont faites pour accroître l'homogénéité du corpus ainsi que des métadonnées.

Mots-clés : numérisation d'archive, corpus hétérogène, classification lexicale, Revolution française.

1. Introduction

La République est née dans le débat, avec ses tribuns célèbres et ses participants plus anonymes : « La "Révolution" aura été ce moment particulier pendant lequel "structure" et "événement" se sont mutuellement transfigurés, provoquant la naissance de nouvelles réalités sociales et discursives, notamment des groupes sociaux inédits, comme les élites d'Etat, ou des identités collectives » (Martin, 2005, p.11).

Nous avons voulu retourner à ces origines en nous attachant à observer les mots des débats fondateurs de l'identité française, de la constitution des groupes et des représentations sociales qui émergeaient. Analyser la Révolution sous l'angle des discours de l'Assemblée nationale a plusieurs conséquences importantes. Cela nous amène d'abord à considérer des thématiques avant la chronologie, ce qui constitue une approche inhabituelle et éventuellement

1 Ce travail a été réalisé dans le cadre du LABEX SMS portant la référence ANR-11-LABX-0066

inconfortable. Car chercher à comprendre la Révolution française, c'est souvent envisager une succession d'événements bien connus, chacun étant établi comme une conséquence du précédent et une cause du suivant. Cela amène aussi à se dégager, au moins provisoirement, de l'emprise des grandes figures de la Révolution. Car les Assemblées étaient pléthoriques : près de 1200 parlementaires en 1789, 742 en 1791 (différents des précédents). Et il n'y a pas de raison de donner plus d'importance à une prise de parole qu'à une autre, quelles que soient la notoriété de l'orateur et l'information dont on dispose sur ce qu'il est, était ou sera. Certains de ces grands noms de la Révolution n'étaient d'ailleurs pas parlementaires tandis qu'ils jouaient un rôle majeur au sein d'instances parallèles et souvent concurrentes. On garde néanmoins leur trace car ils pouvaient intervenir à l'Assemblée.

A l'instar de l'analyse des corpus parlementaires contemporains (Ratinaud & Marchand, 2015), nous nous attachons donc ici à retracer la dynamique de structuration des mondes lexicaux et de leur évolution pour reconstruire la chronologie des débats dans les Assemblées révolutionnaires.

L'objectif de cette recherche est de présenter les premières étapes de la construction d'un corpus dans une situation où les textes disponibles s'éloignent significativement des normes de présentation habituelles dans le champ de l'ADT et pour préparer une analyse plus approfondie à venir.

2. Corpus d'origine

Si l'on se réfère à la bibliothèque interuniversitaire de la Sorbonne[2], « la publication des Archives parlementaires a été inaugurée en 1862 à l'initiative du Corps législatif, sous le Second Empire. D'abord conçue comme une suite de la réimpression du Moniteur universel qui devait rendre plus facilement accessibles les débats de la période 1800-1860, l'édition fut étendue en 1867 à la période révolutionnaire. La publication de la 1ère série fut interrompue par la Première Guerre mondiale, au tome LXXXII, à la séance de la Convention nationale du 15 nivôse an II (4 janvier 1794). En 1956, à l'issue de démarches effectuées par Georges Lefebvre, le CNRS décidait la reprise de la publication, allouant pour ce faire des crédits à l'IHRF-IHMC, qui assure depuis la publication ».

Le corpus repose sur le travail mené en collaboration entre les bibliothèques de l'Université de Stanford et la Bibliothèque nationale de France (BnF), entre 2010 et 2012, pour créer une version numérisée des principales sources d'étude de la Révolution française et les mettre à disposition de la communauté académique internationale[3].

Les ressources textuelles contiennent l'ensemble des délibérations parlementaires de 1789 à 1794 en plus de 82 volumes (pris du tome 8, du 5 mai 1789 au 15 septembre 1789, au tome 82, du 20 décembre 1793 au 4 janvier 1794), constitués chacun d'environ 800 pages qui relatent des évènements de la convocation des États généraux à la période de la Terreur (pour une analyse des *Archives Parlementaires de la Révolution*, voir Gomez-Le Chevanton et Brunel, 2015).

2 http://www.bibliotheque.sorbonne.fr/biu/spip.php?rubrique212

3 https://frda.stanford.edu/fr

Notons que le texte de ces volumes a été codé en format TEI[4] de façon à pouvoir rechercher plus facilement les orateurs, les lieux, les dates et les termes.

Nous avons extrait un corpus portant sur les débats en session de juin 1789 à janvier 1794, que nous avons formaté pour l'adapter à l'analyse par le logiciel *Iramuteq*[5], en codant notamment chaque tour de parole et en repérant le locuteur et la date de la session.

Dans la figure 1, la première colonne représente le texte tel qu'il apparaît dans les ouvrages de l'Assemblée nationale. La partie gauche de cette page contient des annexes au débat ; ces parties n'ont pas été retenues dans notre corpus. La partie de droite présente le début d'une session parlementaire. La colonne du milieu contient un extrait du fichier xml encodé en TEI résulant du travail de numérisation et tagage de l'université de stansford et de la BnF. On y voit, par exemple, la balise <div2 type='session'> que nous avons utilisée pour repérer les sessions. On voit également les balises <date> et <speaker>. La colonne de droite permet de voir le résultat du formatage de ces données pour une analyse par *Iramuteq*. La variable *speaker est ainsi une reprise de la balise <speaker> et nous a servi à délimiter des tours de parole.

Les premières analyses de ce corpus de 43,4 millions de mots environ, montrent que la numérisation est imparfaite. Les statistiques lexicales de base révèlent, par exemple, une fréquence anormale des hapax, c'est-à-dire des formes qui n'apparaissent qu'une seule fois dans le corpus (314582 hapax pour 465 068 formes différentes, soit 67,64% des formes et 0,7% des occurrences). Si l'on compare aux débats à l'assemblée nationale française de 1998 à 2016, ce corpus de 168 millions d'occurrences contient 198754 formes différentes (plus de deux fois moins que le corpus de la révolution) pour 60365 hapax (soit 30,37% des formes et 0,04 % des occurrences).

L'extraction de segments confirme des problèmes de reconnaissance des caractères originaux. Ainsi, le bigramme « lı » peut-il être reconnu comme un h (*répubhque*, *pubhque*, *hberté*, *hvre*...), tandis que « l′ » peut être reconnu comme un V (Vassemblée, Varmée…).

Nous avons donc cherché à éprouver l'hétérogénéité de ce corpus pour voir s'il était pertinent de l'analyser. Pour ce faire, nous avons cherché à accroître le poids des fautes dans un sous-corpus de test soumis à analyse.

4 La « Text Encoding Initiative » (TEI) est une méthode d'encodage de documents textuels définie par une communauté académique internationale dans le champ des humanités numériques pour standardiser un format d'analyse et permettre le partage de ressources et l'échange d'analyses.

5 *Iramuteq* (http://www.iramuteq.org) est un logiciel libre de textométrie, développé par Pierre Ratinaud au sein du Lerass et avec le soutien du Labex SMS

Figure 1: Exemple du passage du corpus original au corpus formaté

3. Sous-corpus de test et recodage des données

Nous avons d'abord extrait de l'index lemmatisé les formes non-reconnues par le dictionnaire d'Iramuteq (codées NR). Nous avons ensuite retiré de la liste les non-reconnaissances qui n'étaient pas problématiques (noms propres sans majuscule, nombres en signes romains, abréviations usuelles…). Nous avons enfin établi la liste suivante par fréquences décroissantes : ae (13815 ; rang 323), dë (4446), ia (4157), ie (4034), dè (3166), ét (3047), â (2371), ët (2276), oe (2201), iv (2007), ies (1998), répubhque (1986 ; rang 1983), ja (1730), lë (1666), èt (1539), vassemblée (1269 ; rang 2859), ge (1253), tit (1252), ar (1203), î (1200), lës (1109), vi lè (1009), qué (998), dp (909), hberté (896), én (865), aes (818), le3 (804), lâ (783), varmée (752), amp (752), cé (725), uu (717), ee (708), pubhc (697), ën (692), quë (688), ur (655), it (647), ui (644), de3 (631), ments (623), el (616), jes (603), res (600), oette (598), is (583), pe (582), nt (580), im (579), di (569), jé (539), ete (529), ést (524), èn (513), tre (497), àu (488), sé (484), dù (475), què (471), së (470), em (469), è (462), ei (452), dës (450), po (439), mo (433), pubhque (406), ront (380), he (380), fe (380), natio (377), tement (375), jë (373), égahté (119).

Ces formes ont été intégrées dans un Type généralisé (Tgen : Lamalle & Salem, 2002). On a donc pu constituer notre corpus-test par extraction de tous les segments de texte qui comportent au moins l'une de ces formes. Ce corpus test comprend 76 197 segments (6,1% du corpus d'origine) et 2 774 518 occurrences (6,4% du corpus d'origine).

Une classification lexicale (Reinert, 1983) permet de décrire la structure du vocabulaire. Sur la base d'une matrice binaire codant la présence (1) ou l'absence (0) d'une forme lexicale dans un segment de texte, on regroupe les segments dont les profils sont semblables dans une arborisation hiérarchique (*dendrogramme*, Figure 1).

La figure 1 montre les corrélations entre les formes lexicales et les classes de la CDH. On y remarque tout d'abord qu'à l'exception de « quë », aucune autre forme non-reconnue ne figure dans les plus fortes corrélations. On remarque ensuite que les formes liberté et hberté, ou égalité et égahté, se retrouvent dans la même classe 6.

Le fait que la classification lexicale s'effectue bien ici sur des segments de texte et non sur les formes lexicales, permet non seulement de neutraliser certaines ambiguïtés, mais aussi d'accepter un certain nombre de fautes graphiques. La répartition des fautes d'océrisation est aléatoire dans le corpus. En conséquence, les fautes se répartissent en proportion de façon équiprobable dans les classes. On peut donc constater qu'aucune de ces formes n'est représentée dans aucune des classes. On en conclut qu'elles n'ont pratiquement pas d'incidence sur l'analyse classificatoire.

Ce résultat nous autorise donc à analyser le corpus d'origine. Mais, dans le but d'accroître l'homogénéité, nous ferons une recherche des concordances des formes non-reconnues de forte fréquence pour les corriger dans le dictionnaire et leur attribuer une catégorie morphosyntaxique. Par exemple :

ae → de (pre) ; 1er → 1er (num) ; ii → ii (num) ; dë → de (pre) ; ia → la (art_def) ; ie → le (art_def) ; xvi → xvi (num) ; dè → de (pre) ; ét → et (con) ; iii → iii (num) ; â → â (pre) ; ët → et (con) ; tion → tion (pre) ; oe → ce → pro_dem) ; iv → iv (num) ; ies → les (art_def) ; répubhque → république (nom) ; ja → la (art_def) ; lë → le (art_def) ; 4e → 4e (num) ; ier → ier (num) ; èt → et (con) ; 2e → 2e (num) ; yous → vous (pro_per) ; vassemblée → assemblée (nom) ; ge → ge (pro_dem) ; tit → tit (pre) ; ar → ar (pre) ; î → î (pre) ; lës → les (art_def)…

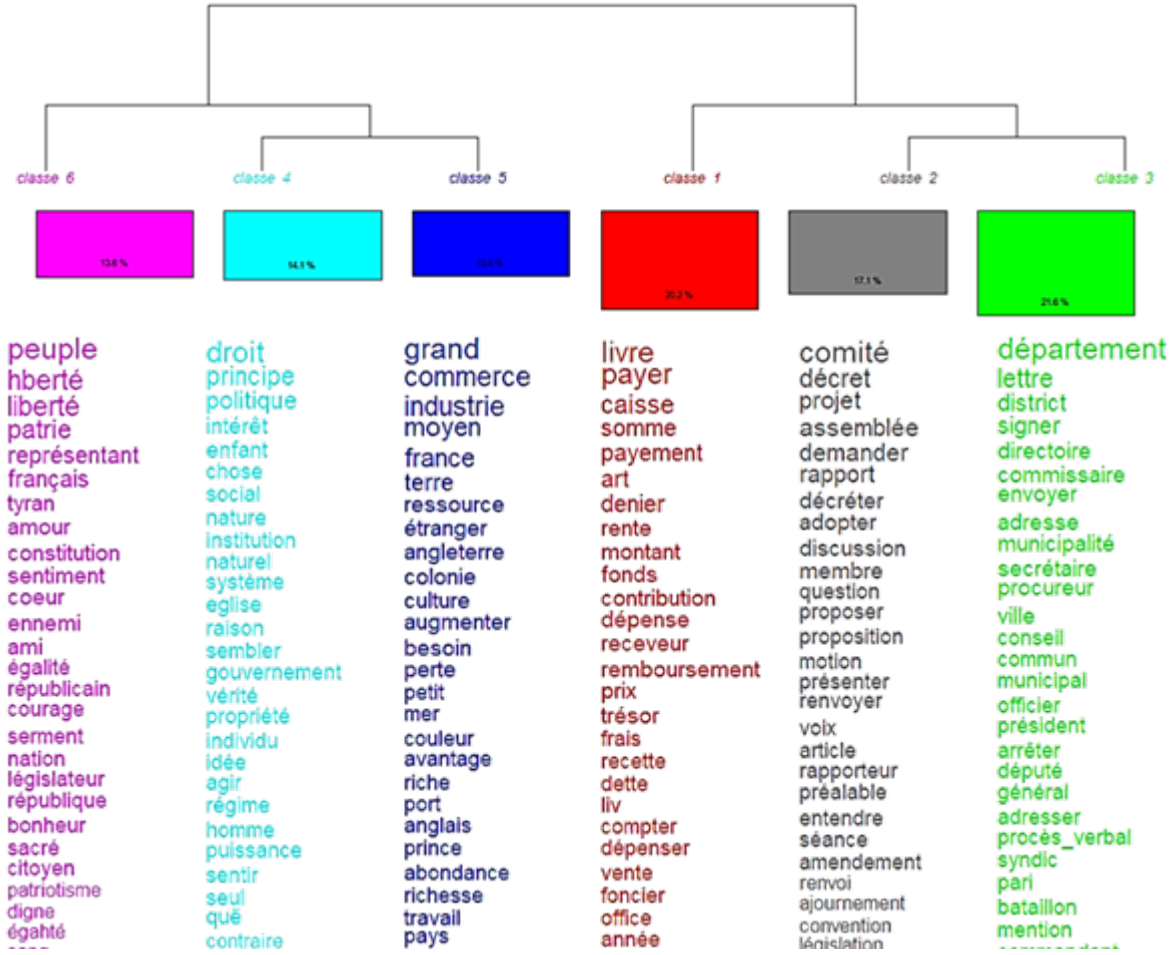


Figure 2: Dendrogramme de la CDH des segments de texte comportant des formes non-reconnues (Méthode Reinert, Iramuteq)

4. Recodage des métadonnées

Deux métadonnées, le locuteur et la date, figurent dans le corpus d'origine et une troisième, le groupe, doit être reconstruite, sachant que le fichier des métadonnées comprend plus de 110 000 lignes.

Les locuteurs sont présents dans le corpus d'origine mais nous retrouvons ici le même problème de reconnaissance des caractères que pour les données textuelles, qui se retrouvent donc dans les balises générées par la numérisation. Il peut s'agir de fautes simplement graphiques (*Albitte / Albilte/ Albitle ; lUaxlinllien…*), de la présence aléatoire de Monsieur ou M., du prénom, du titre (*l'abbé, curé, baron, comte, chevalier, duc, prince*…)[6]. Les mêmes personnes peuvent être nommées de plusieurs façons (Archevêque de Vienne = Lefranc de Pompignan). Il peut y avoir des locuteurs partageant une intervention (*merlindedouai_et_mailhe, merlin_et_basire, thuriot_et_collotdherbois, thuriot_et_jean-bon-saint-andré*) et le locuteur a été recodé en « *loc_multi ».

La désambiguisation manuelle nécessite une comparaison avec des données externes. Certaines homonymies ont pu être résolues avec les dates (Bernard, Bonnet, Fabre, Lameth,

6 Le célèbre Mirabeau doit être distingué de son frère et peut apparaître comme demirabeau, lecomtedemirabeau (juil.1789), mlecomtedemirabeau, levicomtedemirabeau (août 1789-déc 1790), riquetti (1990), riquettidemirabeau (1991)…

Delacroix…). D'autres n'ont pas été résolues et sont actuellement codées comme *loc_??? (Bourdon, Delacroix, Delaunay, Duval, Lindet, Merlin, Prieur, Regnaud, Roux, Varnier…). Le même codage a été affecté à des intervenants qui n'étaient pas députés mais ministres, généraux, membres de comités, personnalités diverses, et auxquels on n'a pas pu attribuer de groupe. Le tableau suivant rapporte les statistiques des locuteurs les plus présents.

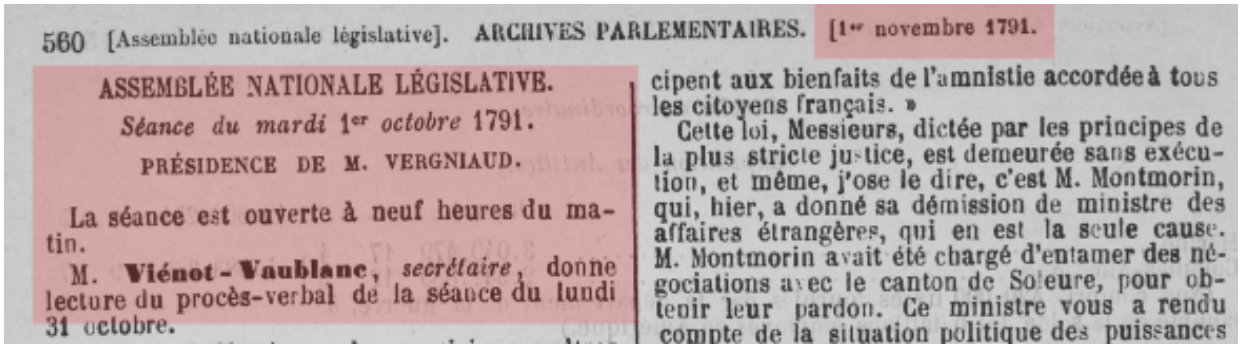
	Constituante	**Législative**	**Convention**	**Total**
*loc_leprésident	5944	6018	3933	15895
*loc_delacroix	6	1414	473	1893
*loc_thuriot	12	831	926	1769
*loc_cambon	10	692	582	1284
*loc_camus	898	14	148	1060
*loc_démeunier	1030			1030
*loc_lanjuinais	584		364	948
*loc_barère	127		807	934
*loc_merlindethionville		673	238	911
*loc_basire		668	231	899
*loc_robespierre	449		438	887
*loc_dandré	857			857
*loc_lechapelier	804			804
*loc_maury	724			724
*loc_malouet	692			692
*loc_marat	1		687	688
*loc_rouyer	5	577	106	688
*loc_lasource	5	440	235	680
*loc_chabot	2	348	317	667
*loc_vergniaud	9	430	226	665
*loc_guadet		440	211	651
*loc_regnaud	650			650
*loc_mirabeau	637			637
*loc_bouche	636			636
*loc_thouret	633		1	634
*loc_defermon	456		159	615
*loc_charlier	5	363	235	603
*loc_prieur-lamarne	599			599
*loc_duport	517	76		593
*loc_martineau	585		5	590
*loc_mathieu-dumas	9	557		566
*loc_decazalès	564			564
*loc_rewbell	482		79	561
*loc_buzot	197		354	551
*loc_barnave	513			513
*loc_bréard		238	271	509
Total	**43744**	**36105**	**30316**	**110165**

Tableau 1 : Locuteurs ayant le plus grand nombre de prises de parole dans les trois assemblées

On observe que la modalité « le président » représente près de 16.000 occurrences. Elle regroupe de multiples locuteurs qu'il faut décrypter manuellement. Ce travail est en cours en utilisant une base de données des 72 présidences des trois assemblées et leurs dates de fonctions. Rappelons que l'Assemblée constituante, par exemple, prévoyait que le président ne pouvait être élu que pour quinze jours consécutifs.

Les fonctions de rapporteur et de secrétaire apparaissent de façon beaucoup moins importante et pourront également être recodées.

En ce qui concerne les dates, présentes en métadonnées sous forme de balises, elles ont simplement été repérées pour permettre l'analyse chronologique des discours selon les années, mois et jours. On a pu, ici encore, repérer des erreurs provenant, soit du fichier d'origine (voir la figure suivante), soit la numérisation (fautes de reconnaissances des caractères numériques).



560 [Assemblée nationale législative]. ARCHIVES PARLEMENTAIRES. [1er novembre 1791.

ASSEMBLÉE NATIONALE LÉGISLATIVE.

Séance du mardi 1er octobre 1791.

PRÉSIDENCE DE M. VERGNIAUD.

La séance est ouverte à neuf heures du matin.

M. **Viénot-Vaublanc**, *secrétaire*, donne lecture du procès-verbal de la séance du lundi 31 octobre.

cipent aux bienfaits de l'amnistie accordée à tous les citoyens français. »

Cette loi, Messieurs, dictée par les principes de la plus stricte justice, est demeurée sans exécution, et même, j'ose le dire, c'est M. Montmorin, qui, hier, a donné sa démission de ministre des affaires étrangères, qui en est la seule cause. M. Montmorin avait été chargé d'entamer des négociations avec le canton de Soleure, pour obtenir leur pardon. Ce ministre vous a rendu compte de la situation politique des puissances

Figure 3: Reproduction d'une page originale avec une erreur manifeste de datation.

Le codage des dates a également permis d'observer la génération de doublons à éliminer.

Enfin, l'attribution des groupes aux locuteurs n'était pas encodée dans le document numérisé et a nécessité une comparaison avec une base externe. Tous les députés des trois législatures sont listés dans la base de données de l'Assemblée nationale française[7]. Cette base a été importée et mise en forme pour la rendre compatible avec notre traitement du corpus. La comparaison des codages du corpus avec la base permettait d'assigner son groupe au locuteur. Mais la condition était que les locuteurs soient écrits, dans le texte numérisé, de façon strictement similaire à la base. On a vu, avec la correction des locuteurs, que ce n'était pas toujours le cas. Par ailleurs, un certain nombre d'absence ou d'erreurs dans l'attribution de groupe (homonymies et anachronismes) ont été constatées sur la base de données de l'Assemblée nationale et ont fait l'objet d'une correction manuelle.

Enfin, la classification des groupes reposant sur la base de l'Assemblée nationale est relativement complexe, comme le montre le tableau suivant. Si les groupes de la Constituante sont assez clairs, les deux autres assemblées connaissent un nombre élevé de groupes qui peuvent éventuellement se recouper. Par ailleurs, les termes « majorité » et « minorité » peuvent renvoyer à des groupes différents avant et après le 2 juin 1793. Enfin, on sait que la dénomination même de ces groupes, mêmes les plus célèbres (« Girondins », par exemple), peut être postérieure à la période révolutionnaire. Une redéfinition et une réduction des groupes par regroupement ne peut se faire sans une parfaite connaissance du contexte et l'aide d'un-e historien-ne sera indispensable à cette étape.

7 http://www2.assemblee-nationale.fr/sycomore/recherche

CONSTITUANTE	***Effectifs***	***pourcentage***
*leg1_non	51280	46.55 %
*leg1_Tiers-Etat	26913	24.43 %
*leg1_???	17790	16.15 %
*leg1_Noblesse	10148	9.21 %
*leg1_Clergé	3832	3.48 %
*leg1_nc	202	0.18 %
LEGISLATIVE		
*leg2_non	52565	47.71 %
*leg2_???	17790	16.15 %
*leg2_Gauche	14549	13.21 %
*leg2_Majoritéréformatrice	8307	7.54 %
*leg2_Modérés	5392	4.89 %
*leg2_Extrêmegauche	3103	2.82 %
*leg2_Droite	1830	1.66 %
*leg2_Majorité	1698	1.54 %
*leg2_nc	1623	1.47 %
*leg2_Constitutionnels	1067	0.97 %
*leg2_Feuillants	645	0.59 %
*leg2_Centredroit	366	0.33 %
*leg2_Plaine	334	0.3 %
*leg2_Minorité	280	0.25 %
*leg2_Girondins	252	0.23 %
*leg2_majoritéréformatrice	184	0.17 %
*leg2_Centregauche	89	0.08 %
*leg2_Centre	42	0.04 %
*leg2_Constitutionnelsmodérés	28	0.03 %
*leg2_Minoritémodérée	17	0.02 %
*leg2_Majorité(Girondin)	4	0 %
CONVENTION		
*leg3_non	44093	40.02 %
*leg3_???	17591	15.97 %
*leg3_Montagne	15470	14.04 %
*leg3_Gauche	15081	13.69 %
*leg3_Girondins	9184	8.34 %
*leg3_Modérés	3794	3.44 %
*leg3_Plaine	1640	1.49 %
*leg3_nc	861	0.78 %
*leg3_Centregauche	670	0.61 %
*leg3_Droite	527	0.48 %
*leg3_Thermidoriens	359	0.33 %
*leg3_Centredroit	225	0.2 %
*leg3_Droitelégitimiste	208	0.19 %
*leg3_Majorité	127	0.12 %
*leg3_Dantonnistes	120	0.11 %
*leg3_Centre	91	0.08 %
*leg3_Gironde	61	0.06 %
*leg3_Minorité	48	0.04 %
*leg3_Marais	12	0.01 %
*leg3_Minoritélibérale	3	0 %

Tableau 2 : liste des groupes des députés des trois premières assemblées (d'après la base de données des députés français en nombre de prises de parole).

5. Première classification du corpus d'origine

Malgré le grand nombre de fautes provoquées par la numérisation, la taille du corpus, la possibilité d'analyser un grand nombre de formes et les résultats du sous-corpus de test nous incitent à en établir les principales thématiques et leur distribution chronologique.

On a gardé les 20 000 formes lexicales les plus fréquentes que l'on a croisées avec 1,2 millions de segments de textes (soit une matrice de 24 milliards de cases).

La CDH permet de définir des classes très claires (Figure 4). Plus précisément, quatre blocs lexicaux se dégagent et renvoient, par exemple, aux procédures des Assemblées, aux considérations économiques et politiques, aux dynamiques identitaires, à la gestion des territoires…

On peut décrire chacune de ces classes par le vocabulaire qui leur est le plus corrélé et en extraire quelques segments caractéristiques. Dans la figure 4, on projette le dendrogramme des 21 classes sur la chronologie (codée en mois sur les cinq années). La hauteur des lignes est proportionnelle à l'importance quantitative des classes (en proportion de segments de texte) et la largeur des colonnes est proportionnelle à la quantité des discours sur la période considérée. Les cases apparaissant en couleur signalent une *surreprésentation* statistique d'une thématique sur une période. Les cases blanches ne signalent pas une absence de ce type de discours, mais une proportion inférieure ou équivalente à la proportion retrouvée dans les autres périodes.

La distribution chronologique des classes (Figure 5) montre clairement des ruptures intervenant à des moments marquants de l'histoire de cette période, que nous avons matérialisé par des lignes verticales. Plus précisément, une première séquence lexicale couvre la période de juillet 1789 à septembre 1791. Une deuxième séquence lexicale s'ouvre en octobre 1791 jusqu'en octobre-novembre 1792, où une troisième séquence court jusqu'en janvier 1794. Ces trois périodes renvoient bien aux trois assemblées fondatrices de la République : *Constituante*, *Législative* et *Conventionnelle*. L'interprétation des thématiques liées à chacune des périodes est en cours.

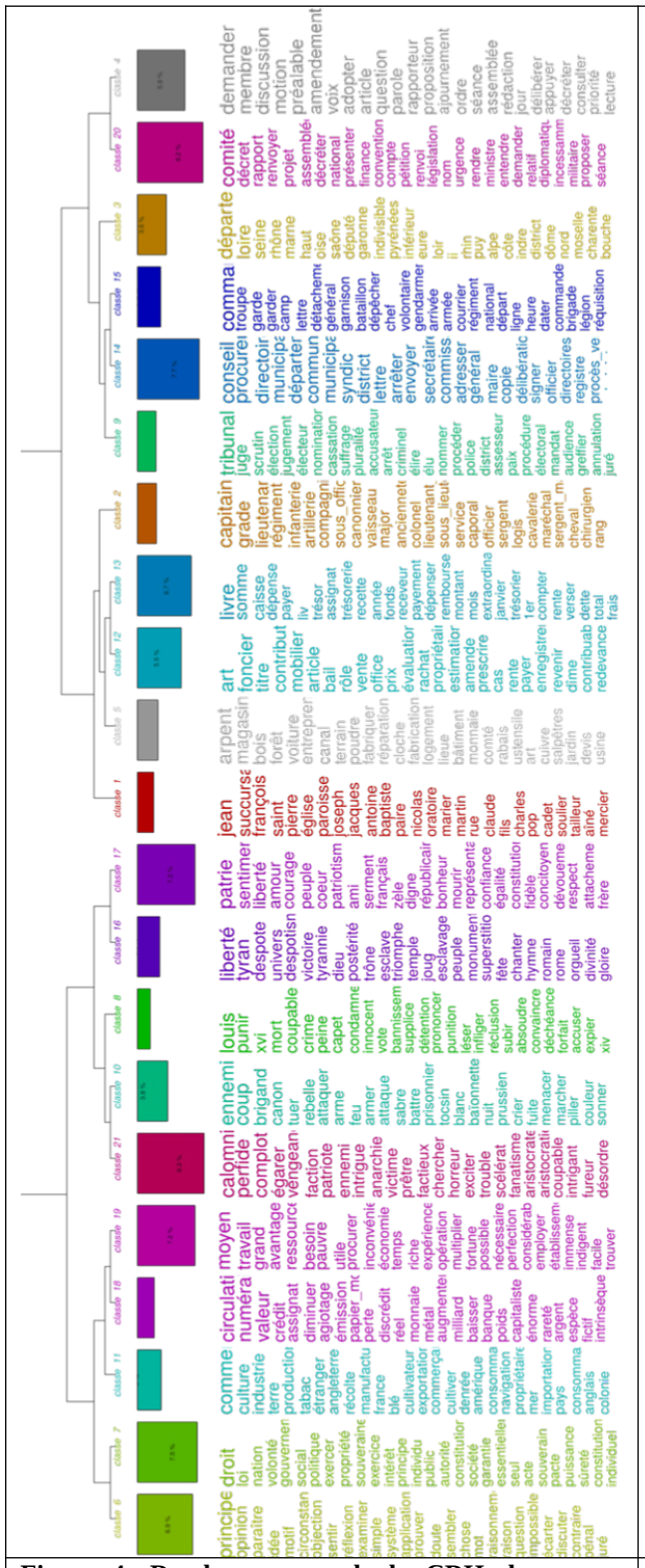


Figure 4: Dendrogramme de la CDH du corpus « assemblées révolutionnaires » (Méthode Reinert, Iramuteq)

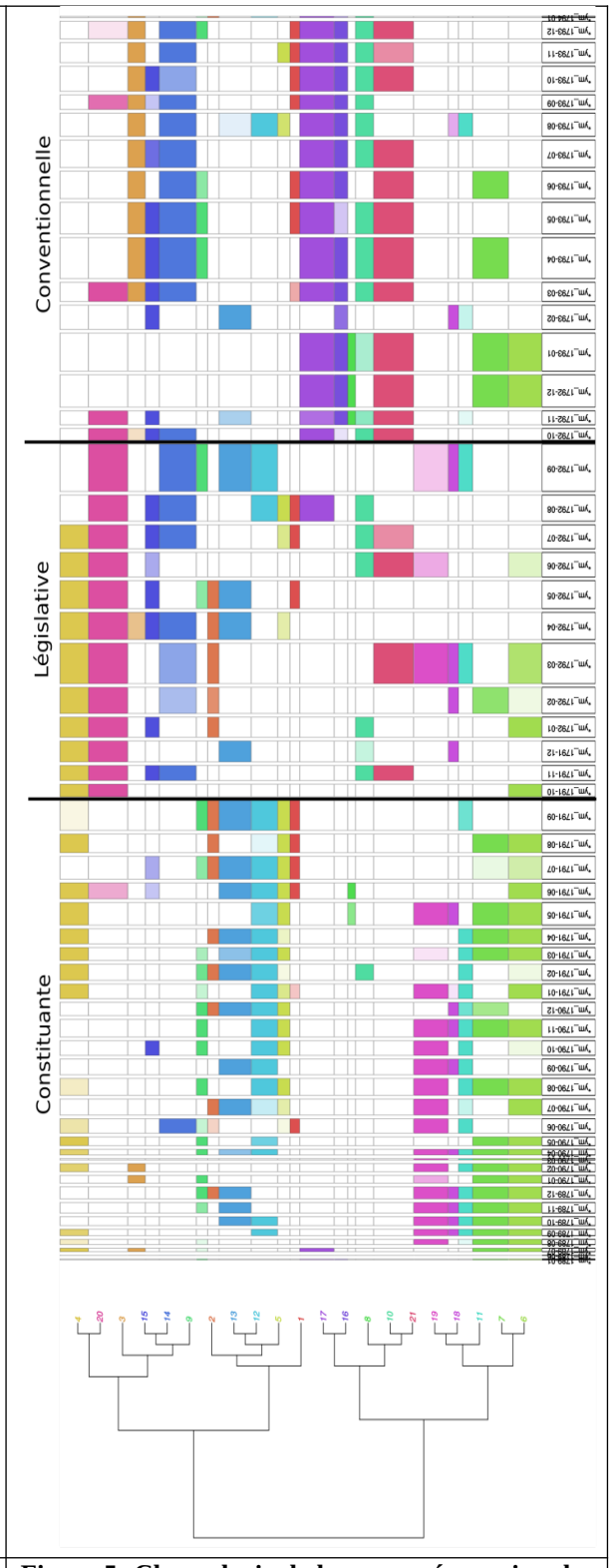


Figure 5: Chronologie de la surreprésentation des classes du corpus « assemblées révolutionnaires »

6. Conclusion

L'objectif du travail présent était de montrer, d'une part comment on peut préparer ce type de corpus massif et hétérogène pour un traitement statistique des données textuelles, et d'autre part que les fautes produites par les processus d'océrisation n'interdisent pas l'analyse effective de ces données. Malgré le soin apporté au codage des données et métadonnées, une masse textuelle aussi considérable ne permettra pas d'atteindre un niveau de précision et d'homogénéité parfois attendu et préconisé (Labbé, 1990) et nous inviterons à la discussion sur la prise en charge de tels corpus.

Il est probable que nos analyses redécouvrent quelques évidences déjà décrites dans une très vaste littérature. On retrouve par exemple des lexiques liés au roi (fuite à Varennes, procès et exécution). On voit se dégager nettement les trois législatures : Constituante, Législative, Conventionnelle, caractérisées par des lexiques spécifiques. Nous aurons néanmoins la satisfaction de penser que la confirmation de ces évidences valide en quelque sorte notre approche.

On voit également apparaître des thématiques plus ou moins étudiées à notre connaissance, comme la réorganisation des territoires ou de l'armée, qui devront faire l'objet d'approfondissement en mobilisant des spécialistes.

Nous n'ignorons pas, enfin, que nos analyses, résultats et interprétations sont susceptibles de nous faire entrer dans des débats, controverses ou polémiques dont nous ne maîtrisons pas totalement les enjeux, courants et références. On montre ainsi que le vocabulaire identitaire et complotiste, justifiant la politique de *Terreur*, n'apparaît que tardivement.

Mais, dans l'état actuel du corpus, nous ne pouvons qu'ébaucher des pistes interprétatives et un travail important sera nécessaire pour progresser.

References

Gomez-Le Chevanton, C. & Brunel, F. (2015). La Convention nationale au miroir des Archives Parlementaires. *Annales historiques de la Révolution française*, 381 (3), 11-29.

Labbé, D. (1990). Normes de saisie et de dépouillement des textes politiques. *Cahier du CERAT*, 7, 1-135. En ligne : https://halshs.archives-ouvertes.fr/file/index/docid/437150/filename/LabbeNormes.pdf

Lamalle, C., Salem, A. (2002). Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels. *Sixièmes journées d'analyse des données textuelles*, Saint-Malo, 403-412.

Martin, J.-C. (dir.) (2005). *La Révolution à l'œuvre. Perspectives actuelles dans l'histoire de la Révolution française*, Presses Universitaires de Rennes.

Ratinaud, P. et Marchand, P. (2015). Des mondes lexicaux aux représentations sociales. Une première approche des thématiques dans les débats à l'Assemblée nationale (1998-2014). *Mots. Les langages du politique*, vol. 2, no 108, 57-77.

Reinert, M. (1983). Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte. *Les Cahiers de l'analyse des données*, 8(2), 187-198.