UNIVERSITÉ PARIS SCIENCES & LETTRES

LE TRAITEMENT DU BOULANGISME PAR LA PRESSE FRANÇAISE

Étude de cas à travers l'OCR du quotidien *Le Matin* au cours de l'année 1887

Directeur de mémoire HN : Florian Cafiero

Directeur de mémoire SHS : Marie Puren

Mini-mémoire de première année du master
« Humanités Numériques »

2022

Résumé

Le boulangisme est un mouvement politique hétéroclite, ralliant de nombreux mécontents au sein d'une Troisième République jeune et encore fragile, et largement critiquée. Le général Georges Boulanger, qui est avant tout un arriviste avec un penchant pour l'autoritarisme, suscite la sympathie voire l'admiration tout autant que la crainte au sein de la population autant que dans les sphères politiques, économiques, et journalistiques, où il suscite autant l'adhésion que le rejet. Le journal *Le Matin* appartient sans aucun doute à la deuxième catégorie, et il s'agit ici d'une première ébauche d'une étude destinée à être élargie, à l'avenir, visant à mesurer la place réservée au traitement du cas Boulanger dans ce journal, ainsi qu'une tentative de mesure de la dynamique populiste impulsée par le général. L'année 1887 est quant à elle riche en rebondissements autour de la personne du général, ce qui en fait un objet intéressant à étudier à l'aide de différents outils relevant de la textométrie ou du traitement automatique des langues (TAL), en nous appuyant sur les textes issus de la reconnaissance optique de caractères (ou *Optical Character Recognition* en anglais) du journal, disponible sur Gallica. Nous y abordons également la nécessité de la mise en place future d'une post-correction de l'OCR qui, s'il reste exploitable, requiert des remaniements.

Mots-clés : Boulanger, presse, Le Matin, OCR, post-correction, TAL, textométrie

Informations bibliographiques : Adrien Escoda, *Le traitement du boulangisme par la presse française : étude de cas à travers l'OCR du quotidien Le Matin au cours de l'année 1887*, mémoire de master 1 « Humanités Numériques », dir. [Florian Cafiero, Marie Puren], Université Paris, Sciences & Lettres, 2022.

Remerciements

Je remercie M. Florian Cafiero d'avoir accepté ma demande de tutorat, malgré ses obligations et le changement de nature de mon projet.

Je remercie également grandement Mme Marie Puren, qui m'a soutenu dans les moments difficiles et a su me redonner confiance et motivation.

Merci à M. Vincent Jolivet pour son aide précieuse sur TXM.

Enfin, je remercie tout particulièrement M. Chahan Vidal-Gorène pour sa disponibilité pour toutes sortes de problèmes que j'ai pu rencontrer (matériel, logiciel, de code), et dont la gentillesse n'a d'égale que sa patience.

Introduction

« On ne saurait trop surveiller ce conspirateur auquel on a eu la faiblesse de laisser un corps d'armée [...]. Il se prépare quelque chose, n'en doutez pas [1] ». Ces propos, tenus par Jules Ferry, ancien président du Conseil de février 1883 à mars 1885, en septembre 1887 dans sa correspondance avec Joseph Reinach, journaliste, républicain modéré et en campagne contre le général Boulanger, atteste bien la crainte que suscite alors l'ex-ministre de la Guerre.

Le boulangisme est un mouvement politique bien difficile à cerner, et même à définir. Est-il même un « mouvement »en 1887 ? Le Comité Républicain de protestation nationale (CPRN, puis CRN par la suite), qui dénonce la mise en non-activité du général prononcée le 15 mars 1888 pour indiscipline, n'est fondé que le lendemain. C'est cet organe qui va alors officiellement et unilatéralement diriger le mouvement « boulangiste »jusqu'à sa dissolution en mai 1890 [2]. Mais en 1887, il n'existe aucune structure pour encadrer ce qui est alors un ensemble hétéroclite de mécontents gravitant autour de la figure du général. Le « boulangisme », c'est d'abord et avant tout la convoitise d'un homme pour le pouvoir. Si mouvement il y a, c'est celui du général vers la conquête des hautes sphères du pouvoir. Le général vise toujours plus haut, mais sans véritable cible précise, quelque soit ce que spéculent ses partisans ou adversaires à ce sujet [3]. Cette ambition n'est pas nouvelle, et l'homme est prêt à tout pour monter chaque échelon, se construit un réseau diversifié, au prix de mensonges et d'alliances diverses, comme celle qui l'unit aux républicains radicaux et lui permet d'accéder en janvier 1886 au poste de Ministre de la Guerre. L'année 1887 ne marque ainsi pas le début de cette « agitation », comme la nomment à plusieurs reprises les rédacteur de l'éditorial politique du *Matin*, mais constitue une phase importante dans la construction de la figure du personnage, et du mouvement qu'il incarne.

On l'a dit, le « boulangisme »est dur à décrire, mais il ne peut se résumer au rôle de « syndics des mécontents [4] ». Cependant, si l'on ne peut le réduire à cette fonction, force est de constater que le général en est à ce moment-là le « rassembleur [5] ». La liste est longue et les raisons multiples. Premièrement, la défiance à l'égard du parlementarisme et ses errances, qui se développent dans toute l'Europe, France y compris. Si la France n'est pas à proprement parler un régime parlementaire, mais plutôt une synthèse entre celui-ci et le régime conventionnel (Chambre unique, soumission totale de l'exécutif à la Chambre), promu par les républicains radicaux (qui s'opposent en cela aux opportunistes), que l'on nomme régime d'assemblée, c'est bien sous cette appellation qu'il est désigné par

1. BN, naf 24877, fol. 266-268, lettre de Jules Ferry à Joseph Reinach, 17 septembre 1887.
2. Bertrand Joly, *Aux Origines du populisme. Histoire du boulangisme (1886-1890)*, Paris, 2022.
3. *Ibid.*, p. 123 et 286-287.
4. *Ibid.*, p. 27.
5. Jean Garrigues, « Boulanger ou la fabrique de l'homme providentiel », *Parlement[s], Revue d'histoire politique*–13 (2010/1), URL : https://www.cairn.info/revue-parlements1-2010-1-page-8.html, p. 15.

ses opposants. L'instabilité ministérielle qui en résulte (ainsi que du mode de scrutin), la faiblesse de l'exécutif (le Président de la République n'a jamais exercé son droit de dissolution de la Chambre des députés, pouvoir issu du compromis entre républicains et monarchistes consacré par les lois constitutionnelles de 1875), mais également l'affairisme et la corruption des parlementaires ou encore le fonctionnarisme, cristallisent la plupart des critiques. La révision du Sénat, considéré par les républicains radicaux comme une « injure »à la souveraineté populaire, est une des grandes revendications, si ce n'est la seule, portée par le général Boulanger, dont le programme politique semble se limiter au triptyque « dissolution-révision-constituante ». Mais les pouvoirs de la Chambre haute sont consacrés par les lois de 1884, ce qui n'a pas pour conséquence d'apaiser les rancoeurs, que le général va instrumentaliser en faisant jouer la dichotomie élite/peuple, la volonté de la première l'emportant sur celle de la seconde. A ce titre, le général semble bien correspondre à la définition du « populisme protestataire »défini par Pierre-André Taguieff, faisant appel à l'anti-élitisme, mais également à la « confiance dans le peuple, méfiance à l'endroit du régime représentatif (antiparlementarisme), appel au peuple par voie de référendum, « hyperpersonnalisation »du mouvement à travers la figure charismatique du leader « viril »et « honnête », discours interclassiste, idéologie de l'unité nationale. [6] ». Le général rassemble moins par ses idées que par les sentiments, y compris celui de la Revanche contre l'Allemagne, ce qui lui vaut d'ailleurs le surnom de « Général Revanche ».

Ainsi, si le général s'est allié aux radicaux en 1885 afin d'accéder au poste de ministre de la Guerre, ce n'est pas parce qu'il partage les mêmes convictions qu'eux (Boulanger est favorable au renforcement de l'exécutif quand les radicaux sont pour son affaiblissement), mais bien par arrivisme, bien qu'il se soit toujours proclamée en faveur de la République (probablement plus consulaire que républicaine dans son esprit). Les radicaux, qui partagent avec lui sa haine du parlementarisme et son projet révisionniste qu'eux-mêmes portent depuis le début des années 1880, sont ainsi alliés à un personnage bientôt encombrant. Peut-on alors qualifier le boulangisme de républicain, ou de bonapartiste ? Est-il un nationalisme avant l'heure ? Nous ne trancherons pas sur ces questions, mais elles ont animé toute une historiographie autour du général et son mouvement. Sans en faire ici tout l'historique, citons l'ouvrage d'Adrien Dansette, *Le Boulangisme* [7], qui le premier réalise une synthèse remarquée sur le boulangisme et son caractères hétéroclite, sans néanmoins se prononcer sur une quelconque nature de celui-ci. Chose que n'hésite pas à faire René Rémond, dans son essai *La Droite en France* [8], qui le présente clairement comme un mouvement bonapartiste. Mais le débat débute réellement dans les années 1970 avec

6. Michel Winock, « Populismes français », *Parlement[s], Revue d'histoire politique*–13 (2010/1), URL : https://www-cairn-info/le-xxe-siecle-ideologique-et-politique--9782262042370-page-209.html, p. 80.

7. Adrien Dansette, *Le Boulangisme*, Paris, 1938, p. 412.

8. René Rémond, *La Droite en France*, Paris, 1954, p. 544.

Zeev Sternhell dans son essai très controversé *La Droite révolutionnaire (1885-1914)*. Les origines françaises du fascisme[9] qui fait du boulangisme un mouvement né à gauche (le nationalisme étant pour lui fondamentalement de gauche), fondant une nouvelle famille à droite, et constituant la première manifestation du fascisme. Si Michel Winock estime que le boulangisme peut être considéré comme un « nationalisme avant la lettre[10] »du fait de son anti-germanisme attesté, il se demande si le boulangisme fut « préfasciste ou prégaulliste[11] ». Pour finir ce tour d'horizon, citons l'ouvrage du canadien William D. Irvine, *The Boulanger affair reconsidered : Royalism, Boulangism and the Origins of the Radical Right in France*[12] qui remet en cause l'idée que le boulangisme serait essentiellement de gauche. Notons également au passage les thèses plus récentes de Jean-Charles Chapuzet[13] (2005) et Mathieu Providence[14] (2007), qui éclairent sous un jour nouveau le mouvement boulangiste, sans oublier bien sûr la remarquable synthèse effectuée par Bertrand Joly sur laquelle nous nous appuyons en grande partie.

Cherchant à nous inscrire dans la continuité de ces travaux, notre projet sur ces deux années de mémoire consistera à analyser le traitement par la presse, et notamment les grands titres (à savoir *Le Journal*, *Le Petit Parisien* et *Le Petit Journal*). Pour l'heure, nous nous intéresserons plus spécifiquement au cas du *Matin*, connu pour sa virulence à l'égard du général. *Le Matin* est fondé en 1884 par des investisseurs américains (John William Mackay et James Gordon Bennett Jr.) désireux de concurrencer le monopole sur les réseaux télégraphiques de la *Western Union* et la suprématie des agences alliées via une série de journaux qui bénéficient directement d'un service télégraphique spécial. Le journal se veut novateur, promouvant un journalisme à « l'américaine », basé sur un journalisme de faits, sans bavardages politiques, « neutre »et « apolitique », avec pour seul but la rentabilité : l'information devient une « marchandise », hautement négociable et commercialisable[15] ». Seule l'éditorial, intégré en première page du journal sur les deux premières colonnes, donne un espace à une expression politique, partagée entre quatre hommes censés représentés les quatre partis dominants (radicaux, opportunistes, royalistes et socialistes).
Le journal s'adresse avant-tout à une certaine élite sociale qui s'intéresse à l'actualité in-

9. Zeev Sternhell, *La Droite révolutionnaire (1885-1914). Les origines françaises du fascisme*, Paris, 1978, p. 441.

10. M. Winock, « Populismes français »...

11. Id., *La Fièvre hexagonale. Les grandes crises politiques (1871-1968)*, Paris, 1986, p. 129.

12. William D. Irvine, *The Boulanger affair reconsidered : Royalism, Boulangism and the Origins of the Radical Right in France*, New-York, 1989, p. 240.

13. Jean-Charles Chapuzet, *Le Général Boulanger et le boulangisme. Des passions politiques à l'oubli (1886-2005)*, Paris, IEP, 2005, 3 vol.

14. Mathieu Providence, *La contribution paradoxale du boulangisme à l'édification de la démocratie parlementaire*, Paris, Paris X, 2007, 2 vol.

15. Dominique Kalifa, Marie-Ève Thérenty et Alain Vaillant, « La civilisation du journal. Histoire culturelle et littéraire de la presse française au XIXe siècle », dans dir. Dominique Kalifa, Philippe Régnier, Marie-Ève Thérenty et Alain Vaillant, 2011, chap. Le quotidien, p. 290.

ternationale et aux marchés financiers, mais fait également place aux évènements français. L'argument principal de vente du journal est son sérieux, sa réactivité et son indépendance, se plaçant au-dessus des luttes partisanes, s'opposant ainsi à la tradition française des journaux d'opinion. Mais celle-ci est déjà depuis longtemps remise en cause par des quotidien comme *Le Petit Journal*, lancé en 1863 et qui se veut lui aussi un quotidien d'information. Mais là où le nouveau journal se démarque des quotidiens de la presse populaire (dont les numéros sont vendus à un sou, soit 5 centimes), outre son prix (10 centimes [16]) et la faiblesse de son tirage (20 000 exemplaires à son lancement, ne dépassant guère les 30 000 exemplaires [17] au cours de la décennie), c'est qu'il cherche à se concentrer sur l'information brute, « pure »produite par des professionnels, privilégiant le fait-divers, le « scoop », le reportage et l'information étrangère, dont il fait sa spécialité. Point de roman-feuilleton comme dans les quotidiens à un sou, et pas de commentaires politiques des faits, expédiés dans un style télégraphiques. Le journal est ainsi essentiellement diffusé à Paris, et est pratiquement inconnu en province, sa diffusion y étant pratiquement supprimée en 1887 [18]. Mais la formule ne prend pas [19], et les ressources du journal sont rapidement épuisées.

Le rédacteur en chef de nationalité anglaise, Alfred Edwards, prend progressivement le contrôle du journal et en devient même le directeur en octobre 1884 [20], affermant l'exploitation du journal l'année suivante [21], prenant alors le contrôle absolu du journal, bien que toujours détenu par des propriétaires américains qui s'en désintéressent de plus en plus, faute de rentabilité. Le nouveau directeur se sert désormais du journal à des fins personnelles, de nature politico-économiques : il se sert du journal pour monnayer son influence [22]. Officiellement apolitique, Alfred Edwards, qui souhaite rester proche du pouvoir, donne une tonalité pro-républicaine, et se veut particulièrement proche des républicains opportunistes et notamment de Maurice Rouvier, nouveau résident du Conseil en mai 1887. Il rend notamment service à ces derniers tout au long de la crise boulangiste en menant une campagne virulente à son égard, service en échange duquel il sera récompensé.

L'année 1887 marque quant à elle la fin de la première partie de l'« boulangiste »(celle du Boulanger ministre) et le début d'un élan de solidarité populaire à son égard qui prélude à la structuration du mouvement l'année suivante. Cette année est marquées par de nombreux rebondissement et affaires ou scandales liées au général, que nous détaillerons dans notre troisième partie, sujets dont *Le Matin* n'hésitent pas à s'emparer.

16. Dominique Pinsolle, *Le Matin (1884-1944). Une presse d'argent et de chantage*, Rennes, 2012, p. 38.
17. *Ibid.*, p. 51.
18. *Ibid.*, p. 52.
19. Id., « *Le Matin*, les affaires et la politique, 1884-1897 », *Le Mouvement social*–232 (2010/3), URL : https://www-cairn-info/revue-le-mouvement-social-2010-3-page-91.html, p. 97.
20. Id., *Le Matin (1884-1944). Une presse d'argent et de chantage*..., p. 46.
21. *Ibid.*, p. 50.
22. *Ibid.*, p. 57.

Ainsi, on comprend mieux la posture du journal à l'égard du général Boulanger, et si nous nous demanderons dans le cadre d'une problématique plus générale dans notre mémoire de M2, quelles sont les différences de traitement du cas Boulanger par la presse, en intégrant notamment les journaux à grands tirages, nous nous intéresserons ici plus modestement à l'importance donnée par *Le Matin* au général Boulanger et aux différentes affaires qui l'entourent au cours de l'année 1887, ainsi que, si possible, à la rhétorique « populiste »de ce dernier.

Pour ce faire nous nous appuierons sur les textes océrisés du journal disponible sur Gallica. Nous nous attacherons dans un premier temps à décrire le processus de reconnaissance optique de caractères, son histoire et son fonctionnement, ainsi que celui plus spécifique de la presse ancienne. Puis nous présenterons notre corpus et notre méthodologie afin d'exploiter celui-ci. Enfin, nous mettrons en avant dans un troisième temps une ébauche des résultats que nous avons obtenus.

Chapitre 1

Contribution à la chaîne de traitement historique par le biais de l'OCR

Notre mémoire s'inscrit dans la chaîne de traitement historique des documents par le biais de l'Optical Character Recognition (OCR), et plus particulièrement celui de la presse. Nous allons ici réaliser un historique des techniques d'OCR et leurs limites, après avoir tout d'abord défini ce qu'est l'OCR, puis nous nous intéresserons plus particulièrement à l'OCR de la presse, en suivant la même démarche.

1.1 Bref survol historique & descriptif des méthodes de reconnaissance de caractères optiques

Nous nous retracerons ici brièvement l'histoire de l'OCR ainsi que son fonctionnement et ses limites.

1.1.1 Brève histoire de l'OCR

Afin de dresser ici un panorama de l'histoire de l'OCR, nous nous appuierons ici essentiellement sur l'ouvrage collectif de Arindam Chaudhuri, Krupa Mandaviya, Pratixa Badelia et Soumya K. Ghosh, *Optical Character Recognition Systems for Different Languages with Soft Computing*[1]

Différentes tentatives de reconnaissance automatiques de caractères ont lieu dès la fin du XIXe siècle et le début du XXe, avec par exemple le développement de dispositifs d'aide aux aveugles par le biais d'expériences avec l'OCR. Mais la version moderne

1. Arindam Chauduri, Krupa Mandaviya, Pratixa Badelia et Soumya K. Ghosh, *Optical Character Recognition Systems for Different Languages with Soft Computing*, Cham, 2017, p. 12-15.

de l'OCR n'est apparue qu'au milieu des années 1940, et prend son essor au début des années 1960 (« OCR de première génération »), avec le développement des ordinateurs numériques et la possibilité d'applications commerciales, se concentrant principalement soit sur des textes imprimés à la machine, soit sur un petit ensemble de textes ou de symboles manuscrits bien différenciés. Au milieu des années 1960 et au débuts des années 1970, les OCR de seconde génération sont apparus (avec notamment la célèbre machine IBM 1287 en 1965), capables de reconnaître les caractères ordinaires imprimés par une machine ainsi que les caractères imprimés à la main, se limitant néanmoins aux chiffres et à quelques lettres et symboles. Au milieu des années 1970, la troisième génération de systèmes d'OCR est apparue, afin de manipuler des documents de mauvaise qualité et de grands jeux de caractères imprimés et écrits à la main. Plus performantes et moins coûteuses (il faut cependant garder à l'esprit qu'avant le milieu des années 1980, seuls quelques milliers de systèmes seulement sont vendus), ces machines ont permis le développement de machines OCR sophistiquées pour les utilisateurs, notamment dans le secteur de la dactylographie, avant que les imprimantes lasers et les ordinateurs personnels commencent à dominer le secteur de la production de textes.

C'est seulement avec la baisse du coût du matériel informatique dans les années 1980 que les systèmes d'OCR ont commencé à être disponibles sous forme de logiciels et plus massivement distribués, permettant un plus haut degré de précision pour la plupart des polices courantes, voire même de reproduire une sortie formatée qui se rapproche du document d'origine, y compris les images, les colonnes et les autres éléments non textuels. Les approches structurelles sont initiées dans de nombreux systèmes, en plus des méthodes statistiques, mais se concentrent essentiellement sur les techniques de reconnaissance des formes sans exploiter les informations sémantiques. Les progrès sont encore plus grands dans les années 1990 avec la mise en place d'outils et méthodologies permettant d'améliorer les performances des systèmes : les méthodes de la reconnaissance de formes sont associées à des méthodologies d'intelligence artificielle. De nos jours, en plus de l'amélioration toujours plus grandes du matériel informatique, miniaturisé et à plus faible coût, les systèmes d'OCR bénéficient également de méthodologies plus complexes et plus performantes telles que les réseaux neuronaux, le modèle de Markov caché ou encore le traitement des langues naturelles, et fonctionnent pour de plus en plus de langues, comme les langues indiennes ou chinoises. Dans les années 2000, l'OCR a été mise à disposition en ligne en tant que service (WebOCR) et même dans des applications mobiles telles que la traduction en temps réel de signes en langue étrangère sur un smartphone, même si ces appareils, qui ne disposent pas d'OCR intégré, doivent passer par une API d'OCR.

1.1.2 Définition, fonctionnement et limites de l'OCR

La reconnaissance de caractère optique est un processus de classification automatique de motifs optiques contenus dans une image numérique. L'identification automatique est le processus par lequel le système de reconnaissance identifie automatiquement des objets, collecte des données à leur sujet et les saisit directement dans des systèmes informatiques, c'est-à-dire sans intervention humaine. Le concept principal de la reconnaissance automatique des motifs est d'abord d'apprendre à la machine quelle classe de motifs peut apparaître et à quoi ils ressemblent. Dans l'OCR, les motifs sont des lettres, des chiffres et certains symboles spéciaux comme les virgules, les points d'interrogation, ainsi que d'autres caractères. L'apprentissage de la machine est effectué en lui montrant des exemples de caractères de toutes les différentes classes. Sur la base de ces exemples la machine construit un prototype ou une description de chaque classe de caractères. Pendant la reconnaissance, les caractères inconnus sont comparés aux descriptions obtenues précédemment et assignés à la classe qui correspond le mieux. Dans la plupart des systèmes commerciaux de reconnaissance de caractères, le processus d'entraînement est effectué à l'avance. Cependant, certains systèmes incluent des fonctions de formation en cas d'inclusion de nouvelles classes de caractères. Il est également parfois nécessaire d'entraîner soi-même un modèle.

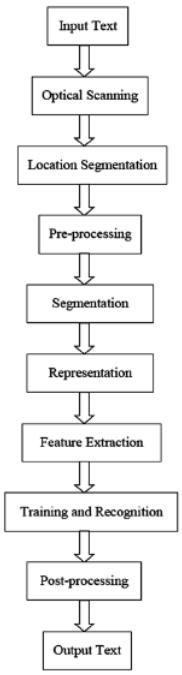


FIGURE 1.1 – Les étapes de la reconnaissance optique de caractères
Source : CHAUDHURI (Arindam), MANDAVIYA (Krupa), BADELIA (Pratixa), GHOSH (Soumya K.), *Optical Character Recognition Systems for Different Languages with Soft Computing*, Cham, Springer International Publishing, 2017, p. 16.

Comme nous pouvons le voir sur le schéma de la figure ci-dessus, plusieurs étapes sont nécessaires avant de pouvoir réaliser un OCR, à commencer par le fait d'avoir de la donnée : c'est tout l'enjeu de la digitalisation des archives anciennes, de tout type (manuscrits, presse, imprimés ...), mise en place dans de nombreux pays dans le monde par diverses institutions. C'est une étape importante : la qualité du document de base et celle de l'image qui en résulte conditionne en grande partie le résultat final. Vient ensuite l'étape de la segmentation, afin de déterminer les constituants d'une image. Il est nécessaire de localiser les régions du document qui contiennent des données imprimées et qui se distinguent des figures et des graphiques. L'étape de *pre-processing*, ou pré-traitement, consiste à d'éliminer le bruit dans une image, de normaliser des données et de compresser la quantité d'informations à conserver, sans perdre d'informations significatives. La binarisation (suppression des échelles de gris) est la première étape importante, car elle permet de réduire la quantité d'informations présentes dans l'image et de n'en conserver que les informations pertinentes. L'élément critique et majeur d'un système d'OCR est la segmentation des lignes de texte à partir des images et consiste à isoler la composante texte dans une image par rapport à l'arrière-plan de l'image. La segmentation des lignes de texte permet ensuite la segmentation des mots puis des caractères. Le principal objectif de la représentation est d'extraire et de sélectionner un ensemble de caractéristiques qui maximise le taux de reconnaissance avec le moins d'éléments possible. L'extraction et la sélection de caractéristiques sont définies comme l'extraction de l'information la plus représentative des données brutes : elles consistent à extraire les caractéristiques pertinentes des objets ou des alphabets pour construire des vecteurs de caractéristiques, afin de permettre la classification, c'est-à-dire l'identification de chaque caractères et son assignation à la classe de caractères correspondante. L'entraînement et la reconnaissance de caractères peut ensuite être effectuée, selon plusieurs approches, soit selon des techniques statistiques, des techniques structurelles, par réseaux neuronaux ou par correspondance de modèles. Enfin, et comme cela est souvent nécessaire, comme dans le cas de notre corpus, une étape de post-traitement doit être effectuée, nous y revenons dans la troisième sous-partie de ce chapitre.

Ainsi on peut voir qu'à chaque étape, la reconnaissance automatique de caractères peut être compromise : mauvaise qualité des documents originaux, numérisation d'images bruyantes et/ou à basse résolution, pré-traitement de l'image incorrect ou insuffisant, mauvaise segmentation des lignes ou des caractères, difficultés liées à la ponctuation, à la typographie etc. La méthode de reconnaissance des caractères elle-même peut ne pas répondre correctement à l'ensemble des caractères donnés, ce qui entraîne des erreurs de classification. Le perfectionnement des techniques et méthodes d'OCR peuvent ou pourront peut-être permettre à l'avenir de surpasser ces difficultés (meilleur traitement initial des images, classification adaptative de caractères, ou encore utilisation accrue du contexte dépendant des propriétés linguistiques du document), mais, dans certains cas,

comme celui de la presse, les marges d'amélioration sont limitées.

1.2 Etat de la numérisation et de l'OCR de la presse ancienne

Nous dresserons ici un panorama de l'état de la numérisation et l'OCR de la presse ainsi que les contraintes liées à cette dernière.

1.2.1 Panorama des différents projets liés à la numérisation et l'océrisation de la presse

La numérisation de la presse ancienne (soit la presse antérieure à 1944 en ce qui concerne la France) est la première étape nécessaire avant son océrisation. Mise en place essentiellement pour les archives de plus de 70 ans, tombées dans le domaine public, elle se développe surtout au carrefour de la fin des années 1990 et du début des années 2000. La Bibliothèque Nationale de France met en place un programme dès 2004 (mis en oeuvre l'année suivante), dont 1200 titres océrisés entre 2007 et 2010 [2]. Les entreprises journalistiques elles-mêmes numérisent leurs archives, à l'instar de celles du *Monde* [3]. Des projets plus globaux sont mis en place, comme l'*Europeana Newspapers Project*, mis en place par la Commission européenne dans son programme 2007-2013, et ayant débuté en 2015 [4]. Ce projet rassemble plus de 21 millions de pages numérisées dont 12 millions sont océrisées, dans 16 langues différentes, couvrant plus de 3400 titres [5]. L'*International Coalition on Newspapers* estime quant à lui dans son rapport du 15 avril 2015 que le nombre de titres numérisés (même s'il ne s'agit pas nécessairement de collections complètes) excède 30 000, voire se rapproche de 45 000 à l'échelle mondiale [6]. On peut retrouver la liste de ces collections à cette adresse : http://icon.crl.edu/digitization.php.

Le processus d'océrisation de la presse numérisée apparaît plus tardivement, à la fin des années 2000, notamment avec le développement du logiciel ABBYY Finereader, qui a permis à la technologie d'OCR d'obtenir de meilleurs résultats pour les journaux historiques. Plusieurs projets ont été développés, outre l'*Europeana Newspapers Project*, à l'instar du projet IMPACT mis en place par la Commission européenne et ayant débuté en 2013,

2. https://pro.bpi.fr/la-numerisation-de-la-presse/, mis en ligne le 15/09/2014, consulté le 26/05/2022.
3. https://www.lemonde.fr/archives-du-monde/
4. http://www.europeana-newspapers.eu/, consulté le 26/05/2022.
5. http ://europeananewspapers.github.io/ ?page=1, consulté le 26 mai 2022.
6. *The "State of the Art" : A Comparative Analysis of Newspaper Digitization to Date*, rapp. tech., ICON, 2015, URL : https://www.crl.edu/sites/default/files/d6/attachments/events/ICON_Report-State_of_Digitization_final.pdf, consulté le 26 mai 2022.

rassemblant notamment la British Library et la Bibliothèque nationale de France, rassemblant plus d'un million et demi d'images textuelles de 1500 à nos jours, bien que la grande majorité porte sur le XIXe siècle [7]. Le projet a notamment pour but de promouvoir la validation et l'adoption d'outils et de ressources qui améliorent l'accès au contenu textuel numérisé et contribuer à la normalisation et à la standardisation des techniques et des résultats de la numérisation des textes [8], et on participé au développement d'outils et logiciels comme Aletheia, qui permet d'océriser des documents, mais également d'enrichir les documents océrisés par le « vérité terrain »(ou *ground truth*), sur laquelle nous reviendrons dans notre deuxième partie. Citons également pour la France le projet Retronews, le site de presse de la Bibliothèque nationale de France, qui donne accès librement et gratuitement à plus de 2000 titres de presse publiés entre 1631 et 1951, qui ont fait l'objet d'une retranscription du texte par OCR et de traitements sémantiques permettant des recherches approfondies et ciblées dans les archives. Ce projet est le complémentaire des archives disponibles sur Gallica, qui propose également des version océrisées, l'une brute, et l'autre corrigée [9], faisant l'objet de corrections manuelles (à l'exception de certaines zones comme les encarts publicitaires dans le cas de la presse).

1.2.2 Facteurs affectant l'OCR de la presse

Malgré toutes les précautions prises au niveau du pré-traitement et lors de l'océrisation à proprement parler, il demeure des limites plus ou moins dépassables qui affectent le résultat final, à commencer par la dégradation des images lors de la numérisation de journaux historiques (en plus de la qualité souvent mauvaise du papier, comme c'est particulièrement le cas pour la presse du XIXe : texte maculé, délavé et déformé ; originaux déchirés ou froissés ; image débordant du verso du papier ; lignes de texte tordues et incurvées ; scans de microfilms surexposés ou sous-exposés. Le contenu de l'image en elle-même peut entraîner des erreurs voire l'impossibilité pure et simple de la retranscription : polices incohérentes/mélangées texte estompé, caractères manquants, traversée de la page - dans certains cas, l'encre de la page opposée "traverse" la page, ce qui rend difficile la distinction du texte par le logiciel.
Les erreurs d'OCR font partie d'un problème plus large de traitement du « bruit »dans l'exploration de texte, qui peut également provenir d'autres sources telles que les variations orthographiques historiques ou le langage spécifique à différents textes. Le processus de segmentation des pages peut également affecter le traitement de l'image en provoquant notamment des confusion de colonnes, notamment pour les pages obliques ou à la mise

7. https://www.primaresearch.org/www/media/datasets/ImpactRepositoryPoster.pdf, consulté le 26 mai 2022.
8. https://www.succeed-project.eu/, consulté le 26 mai 2022.
9. https://api.bnf.fr/fr/ocr-corrige-de-documents-de-presse-de-gallica

en page complexe. Enfin, les problèmes peuvent venir des logiciels eux-mêmes : problèmes liés aux images modèles de la base de données du logiciel, l'algorithme lui-même, ou encore les dictionnaires intégrés. Cependant, l'impact des erreurs d'OCR varie également beaucoup en fonction de la tâche effectuée. Les tâches de détection des limites de phrases, de tokénisation et d'étiquetage de la parole partielle sur le texte sont toutes compromises par les erreurs d'OCR [10].

1.3 La nécessité de l'étape du post-traitement

Le *post-processing*, ou post-traitement, est ainsi une étape nécessaire afin de détecter, regrouper et corriger les erreurs durant le processus d'océrisation. Dans les problèmes avancés de reconnaissance optique de texte, un système ne comprenant que la reconnaissance d'un seul caractère n'est pas suffisant, et même les meilleurs systèmes de reconnaissance systèmes de reconnaissance ne donneront pas une identification correcte à 100%. Il existe deux types d'erreurs d'OCR [11] : les « *non-word errors* »et les « *real word errors* ». La première désigne mot qui est reconnu par le système OCR, mais qui ne correspond à aucune entrée du lexique. A l'inverse, le deuxième type d'erreur correspond à des mots reconnu par le système OCR et qui sont conforment à une entrée du lexique, bien qu'il soit grammaticalement incorrect par rapport à la phrase en question. Par exemple, pour un texte tel que « How is your day ? », avec pour sortie « Huw is you day ? », « Huw »est une erreur de non-mot et « you »une erreur de mot. En général, les erreurs de mots et de non-mots relèvent de trois catégories d'erreurs : les erreurs de suppression, les erreurs d'insertion et les erreurs de substitution. L'erreur de suppression se produit lorsqu'un ou plusieurs caractères sont supprimés ou retirés à l'intérieur du mot original ; l'erreur d'insertion, lorsqu'un ou caractères supplémentaires sont ajoutés au mot original ; l'erreur de substitution, lorsqu'un ou plusieurs caractères sont accidentellement modifiés dans le mot original. Par conséquent, d'innombrables approches et algorithmes de post-traitement et d'algorithmes ont été proposés pour tenter de détecter et de corriger les erreurs d'OCR. Nous citerons ici les les trois approches les plus courantes : la correction manuelle, la correction basée sur le contexte et la correction basée sur un dictionnaire.

La façon la plus simple de corriger les erreurs d'OCR est d'engager un groupe de personnes pour s'asseoir et essayer d'éditer manuellement le texte de sortie de l'OCR, en comparant

10. Daniel McNamara, Strange Carolyn, Josh Wodak et Ian D. Wood, « Mining for the Meanings of a Murder : The Impact of OCR Quality on the Use of Digitized Historical Newspapers. » *Digital Humanities Quarterly*, vol. 8–1 (2014), URL : https://www.dlib.org/dlib/march09/holley/03holley.html.

11. Mohammad Awani et Youssef Bassil, « OCR post-processing error correction algorithm using google's online spelling suggestion », *Journal of Emerging Trends in Computing and Information Sciences*, vol. 3–1 (2012), URL : http://www.cisjournal.org/journalofcomputing/archive/vol3no1/vol3no1_7.pdf, consultée le 28 mai 2022.

des documents numérisés et leurs sorties océrisées par exemple. Mais cette tâche nécessite un grand nombre de personnes à mobiliser, et est coûteuse, chronophage et laborieuse, en plus d'être possiblement sujette à de nombreuses erreurs malgré les relectures successives. La correction basée sur le contexte, bien plus complexe à mettre en place, permet de détecter et corriger des erreurs sur la base du contexte grammatical et parfois du contexte sémantique -ce que ne fait pas la correction basée sur un dictionnaire-, et fonctionnent à partir de modèles de langages statistiques ou des méthodes basées sur les caractéristiques des mots. Il s'agit par exemple d'appliquer un marqueur de « *part-of-speech* »(autrement dit, d'assigner aux mots une catégorie grammaticale : verbe, adverbe, noms, entités nommées etc.) et les règles grammaticales de la langue en question pour capturer les erreurs d'interprétation (en indiquant par exemple, pour le cas de la langue anglaise, qu'un verbe à la troisième personne du présent doit toujours prendre un « s »). La correction basée sur le contexte peut être combinée à une approche basé sur un dictionnaire, celle-ci servant à remplacer les non-mots tandis que la première corrige les erreurs de mots en fonction de leur contexte.

La correction manuelle étant inenvisageable à notre échelle et la correction basée sur le contexte trop complexe à mettre en place, nous privilégions l'approche basée sur un dictionnaire dans le cadre de notre M2. Également connue sous le nom de « correction d'erreurs lexicales », elle consiste à utiliser un lexique ou un dictionnaire de consultation afin de vérifier l'orthographe des mots reconnus par l'OCR et les corriger si ils sont mal orthographiés, en établissant éventuellement une liste de plusieurs mots possibles. Plusieurs méthodes existent, que l'on peut regrouper en deux catégories : les méthodes syntaxiques (basées sur la structure grammaticale d'un mot en prenant en compte son contexte dans la phrase) et les méthodes sémantiques (basées sur le sens des mots). Parmi les méthodes syntaxiques, l'une d'entre elles consiste par exemple en un algorithme de correspondance des chaînes de caractères qui pondère les mots dans un texte en utilisant une métrique de distance représentant divers coûts. Le candidat à la correction ayant la distance la plus faible par rapport au mot mal orthographié est la le mieux adapté à la correction. L'approche sémantique peut quant à elle consister en l'étiquetage des documents dans lequel chaque document numérisé est classé sémantiquement selon son sujet en utilisant un modèle de formation non supervisé. Chaque mot mal orthographié est ensuite corrigé en sélectionnant le candidat à la correction qui appartient à la même classe que l'erreur réelle. On peut également combiner les deux approches en analysant des phrases afin de déterminer si elles sont syntaxiquement et sémantiquement correctes. Lorsqu'une phrase semble suspecte, des candidats possibles à la correction sont générés à partir d'un dictionnaire et regroupés de haut en bas en fonction de leur force syntaxique. A long terme, le candidat en tête de chaque groupe est celui qui remplace l'erreur d'OCR correspondante. Des modèles plus sophistiqués utilisant le modèle de Markov caché ou

le *machine learning* (auto-identification et auto-correction d'erreurs) existent également, plus performants, mais également plus complexes à mettre en place.

Cette approche souffre néanmoins de plusieurs limites. La première est la nécessité de concevoir un dictionnaire étendu qui couvre chaque mot de la langue, si possible en adéquation avec le vocabulaire spécifique d'une époque donnée. Dans notre cas, il s'agira de compiler des mots venant de contextes et sources divers du XIXe siècle : nous nous appuierons sur le paquet *Python pyspellchecker*[12] pour créer ce dictionnaire, en le complétant à l'aide des OCR corrigées de la presse[13] disponibles sur les API de la BNF, ainsi que des textes d'autres domaines comme la littérature[14] et juridiques ou politiques (en utilisant *Wikisource*). La deuxième limite est la prise en charge simultanée de plusieurs langues simultanément (ce qui n'est pas un problème très important dans notre cas), les dictionnaires visant généralement une langue spécifique. La troisième limite réside dans le fait que les dictionnaires conventionnels ne ne prennent pas en charge les noms propres et personnels, les noms de pays, de régions, de lieux géographiques et de sites historiques, les mots-clés techniques, les termes spécifiques à un domaine, les acronymes ou encore des expressions argotiques, et l'enjeu de notre mémoire sera de tenter de pallier au moins en partie ces biais. La dernière limite, mais qui là encore ne concerne pas notre corpus, est le caractère statique d'un dictionnaire qui doit être mis à jour continuellement afin d'intégrer l'évolution de la langue.

12. https://pyspellchecker.readthedocs.io/en/latest/
13. 9
14. github.com/COST-ELTec/ELTeC-fra

Chapitre 2

Le corpus : la version océrisée du *Matin* de l'année 1887

Nous allons présenter dans cette partie notre corpus ainsi que la méthodologie utiliser afin de le traiter et les différentes pistes à mettre en oeuvre pour réaliser ce traitement, à savoir une ré-océrisation (des images) par le biais de Tesseract et la comparaison des « vérités terrain »de l'OCR de Gallica et celui de Tesseract.

2.1 Présentation et récupération du corpus

Nous veillerons dans cette sous-partie à présenter notre corpus, la source et manière par laquelle nous l'avons récupérer.

2.1.1 Le corpus : les quotidiens du *Matin* de l'année 1887

2.1.2 Présentation du matériau original

Nous nous appuierons sur le corpus des deux premières pages de chaque quotidien de l'année 1887 du *Matin*. Comme nous l'avons mentionné en introduction, cette année est riche en rebondissement concernant le cas Boulanger, et se situe également durant la période de réorganisation du journal (entre 1885 et 1890) par son directeur, Alfred Edwards. Celui-ci est à la fois directeur du journal et directeur en chef de la rédaction, et a également placé ses hommes à la tête du conseil d'administration de la maison mère (américaine) détentrice du journal [1], et possède par conséquence une mainmise quasi totale sur le contenu du quotidien. Outre sa volonté d'accélérer les ventes du journal pour le rendre rentable, incitant celui-ci à rebondir sur le moindre faits-divers qui puisse vendre, les intérêts de son chef avec le monde politique donne une tonalité particulière au traitement des affaires liées au général.

1. D. Pinsolle, *Le Matin (1884-1944). Une presse d'argent et de chantage*..., p. 49.

La raison pour laquelle nous ne nous intéressons qu'aux seules deux premières pages tient à la manière dont est constitué le journal, que nous allons brièvement présenter ici.

Quatrième année. — N° 1,243 Paris et Départements : 10 centimes Mercredi 13 Juillet 1887

Le Matin

DERNIERS TÉLÉGRAMMES DE LA NUIT

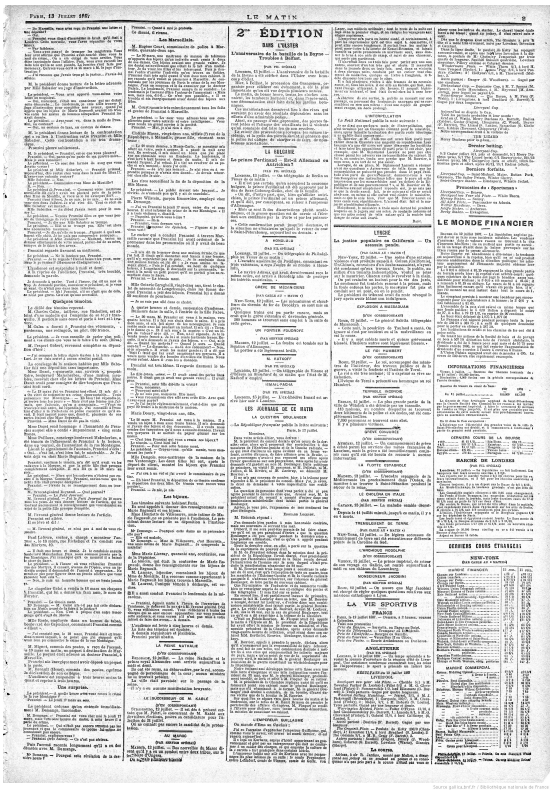
SEUL JOURNAL FRANÇAIS RECEVANT PAR FILS ET SERVICES SPÉCIAUX LES DERNIERES NOUVELLES DU MONDE ENTIER

LA FÊTE DU RÉGIMENT

LE GÉNÉRAL MIRLITON

PRANZINI

FIGURE 2.1 – La première page du quotidien du 13 juillet 1887



2e ÉDITION

FIGURE 2.2 – La deuxième page du quotidien du 13 juillet 1887

Le journal est un grand format (environ 45x60cm), ce qui permet de multiplier les articles et les rubriques d'information, répondant à la mission première que s'est donné le journal. La première page est composée d'un encart de titre, avec en dessous un petit encart avec la devise du journal « Seul journal français recevant par fils et services spéciaux les dernières nouvelles du monde entier », appuyant ses principaux arguments de vente : la réactivité et rapidité dans l'obtention des dernières informations grâce à ses réseaux télégraphiques spéciaux (le journal est branché sur les fils spéciaux du *Morning News* dont *Le Matin* est l'édition française, mais qui en devient indépendant en avril 1885 [2]) et les services de l'agence de presse Havas, dont le journal devient un des plus gros clients à la fin de l'année 1884) ; et son ouverture à l'actualité étrangère auquel le journal accorde une importance particulière et plus poussée que ses contemporains. La page est ensuite divisée en cinq colonnes, elles-mêmes plus ou moins divisées en sections plus ou moins permanentes. La première, à cheval sur la première et la deuxième colonne, comporte l'éditorial politique, dont on a déjà dit en introduction qu'il constituait le seul espace où un commentaire

2. *Ibid.*

politique est accepté, le journal prétendant fournir une information « pure », débarrasser de toute opinion politique, dans le reste des articles [3]. On peut également constater avec la figure 2.1 l'importance de l'article choisi en troisième colonne, qui attire immédiatement l'oeil du fait de sa position centrale, et donc de l'importance du choix éditorial de l'article situé à cette place. Les colonnes s'inscrivent dans les pages suivantes sur toute la longueur de la page, comme on peut le constater avec la figure 2.2. Cette manière de regrouper les articles en colonnes d'y ajouter un titre n'est pas nouvelles et se généralise dès les années 1840-1850 [4] mais le formatage rubrical s'intensifie dans les années 1870-1880 avec le développement de rubriques-titres au contenu fluctuant. *Le Matin* propose une mise en page spectaculaire en généralisant la pratique de l'ajout d'un menu décrivant le contenu de l'article.

FIGURE 2.3 – La troisième page du quotidien du 13 juillet 1887

FIGURE 2.4 – La quatrième et dernière page du quotidien du 13 juillet 1887

On peut ici constater que le contenu des troisième et quatrième page est essentiellement composé de potins sur la vie mondaine, d'articles sur les dernières décisions de justice, des cours de bourses et autres dépêches financières et commerciales, des programmation de spectacles et de publicités. Si la rubrique des dernières informations concernant le

3. *Ibid.*, p. 38.

4. D. Kalifa et M.È. Thérenty, « La civilisation du journal. Histoire culturelle et littéraire de la presse française au XIXe siècle », dans dir. Dominique Kalifa, Philippe Régnier, Marie-Ève Thérenty et Alain Vaillant, 2011, chap. Ordonner l'information, p. 883-885.

Parlement ou le conseil des ministres n'est pas inintéressante pour notre objet d'étude, nous avons fait le choix de laisser de côté ces deux dernières pages qui ont un contenu informatif qui nous a semblé peu pertinent pour notre analyse et qui sont souvent très bruitées, particulièrement la quatrième page avec toutes ses publicités (même si ce n'est pas particulièrement flagrant dans l'exemple ci-dessus).

2.1.3 Disponibilité et méthodes de récupération du corpus

Le texte océrisé et les images des pages sont intégralement disponibles sur Gallica [5]. Concernant l'OCR, il s'agit ici du OCR brut qui ne fait pas l'objet de post-traitement, contrairement à d'autres jeu de presse corrigé, et dont le taux de reconnaissance de caractères est généralement estimé à plus de 98% (soit deux mots sur cent potentiellement erronés), en excluant du calcul les parties les plus illisibles du document [6], ce dont nous pouvons douter au vue des sorties texte que nous avons récupérées (un exemple est fournie dans la deuxième sous-partie). Les images bénéficient d'un pré-traitement (redressement de la page, suppression des bords ; binarisation : conversion en noir et blanc ; segmentation : extraction des zones de la page, découpage en lignes, mots, caractères [7]) ainsi que d'un contrôle de la qualité de la segmentation et de la reconnaissance des contenus textuels de la page. Quant aux images, elles sont récupérables (ainsi que leurs métadonnées) via l'interface de programmation d'application (API) de récupération des images de Gallica [8], en haute définition et aux normes IIIF (International Image Interoperability Framework), qui est une méthode standardisée de description et d'accès aux images sur le Web.

Nous avons utilisé l'outil Pyllica [9] afin de récupérer le texte océrisé, et c'est également par ce biais que nous pourrons récupérer les images pour la ré-océrisation, de manière automatisée. Il s'agit d'un outil écrit en *Python* spécifiquement conçu pour de récupérer des documents hébergés sur la bibliothèque numérique Gallica, et notamment les périodiques de la presse ancienne grâce à ses deux outils consacrés, pyllicalabs pour le texte océrisé et pyllicalabsjpgpress pour les images, aisément accessibles sur github [10]. Il suffit d'indiquer l'url de la page du périodique recensant toutes les années disponibles (en retirant la fin de l'adresse « /date »pour pyllicalabsjpgpress), le titre, accolé au nom du fichier qui sera extrait (ici, « lematin », la date du premier numéro à extraire (l'année, le mois puis le jour, séparément ; en l'occurrence pour nous : 1887, 01, 01), le nombre de fichiers que l'on veut récupérer avec la commande « *item* »(365 dans notre cas), le nombre de jour entre chaque numéro avec la commande « *rate* »(ici, 1), et le numéro de la dernière page que

5. https://gallica.bnf.fr/ark:/12148/cb328123058/date1887
6. https://gallica.bnf.fr/edit/und/consulter-les-documents
7. https://www.bnf.fr/fr/techniques-et-formats-de-conversion-en-mode-texte
8. https://api.bnf.fr/fr/api-iiif-de-recuperation-des-images-de-gallica
9. https://api.bnf.fr/fr/extracteur-python-de-corpus-de-periodiques
10. https://github.com/Dorialexander/Pyllica

l'on veut extraire avec la commande « *lastpage* »(2 en ce qui nous concerne).

2.2 Méthodologie

Détaillons ici notre méthodologie à l'aide d'un schéma qui nous permettra également d'identifier les étapes de travail que nous avons effectuées (cases en vert) et celles qui restent à réaliser (cases en bleu).

2.2.1 Présentation générale des différentes étapes

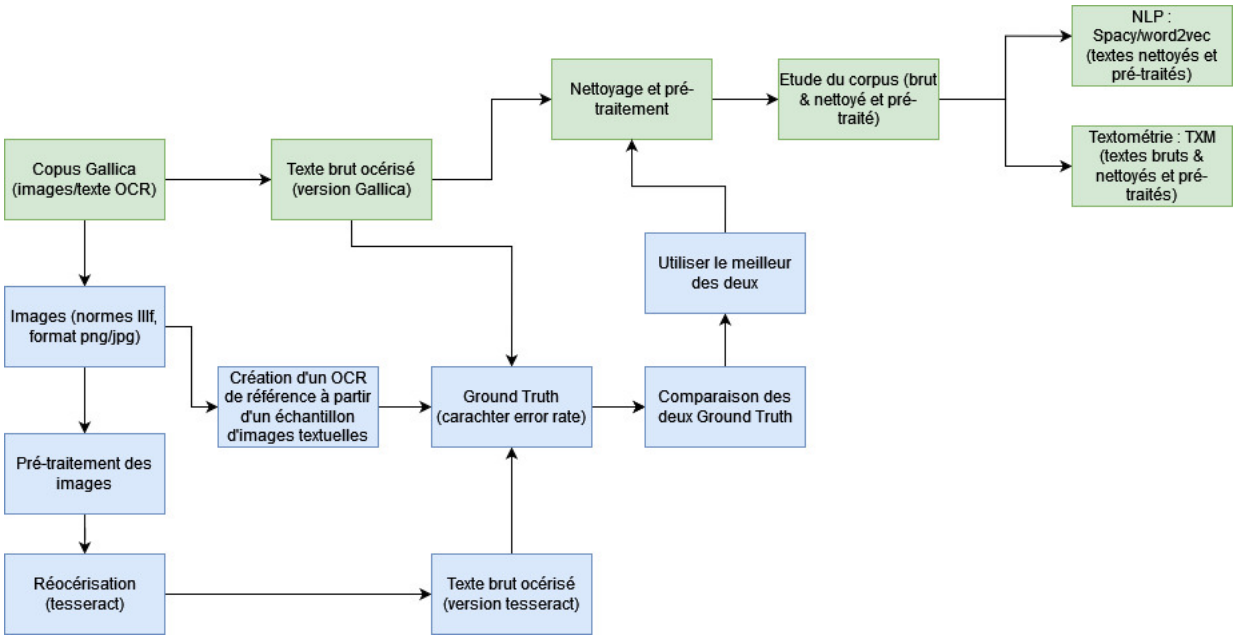


FIGURE 2.5 – Schéma de la méthodologie adoptée afin de traiter notre corpus

Nous avons déjà évoqué la première étape consistant à récupérer le texte ou les images. Nous n'avons ici cependant pas effectuée la deuxième opération à cause de contraintes matérielles (ordinateur et serveurs de Gallica défectueux) et temporelles. La partie en vert correspond aux tâches que nous avons effectuées, comme le nettoyage et le pré-traitement du texte et l'étude du corpus selon deux méthodes distinctes : le traitement automatique des langues (TAL, ou *Natural Processing language* (NLP) en anglais). Nous y reviendrons plus en détails dans notre troisième partie.
Si nous n'avons pas pu effectuer de ré-océrisation par le biais de Tesseract, notre projet dans le cadre de ce M2 serait d'en réaliser une, afin de déterminer quelle est la meilleure version à utiliser. Nous détaillons les raisons de la ré-océrisation dans la sous-section suivante. Il s'agit dans un premier temps de pré-traiter l'image : remise à l'échelle si nécessaire (en 300 ppi), élimination des bordures du texte, suppression du bruit, et binarisation des images. Autant d'opérations qu'il est possible de réaliser avec l'outil Aletheia

que nous avons déjà décrit [11]. Si la segmentation est, comme nous l'avons déjà dit, une étape cruciale dans le pré-traitement de l'image, Tesseract fournit plusieurs modes de segmentation automatisés [12] qu'il s'agira de tester pour déterminer le plus pertinent. Une fois le texte océrisé obtenu, il faudra alors le comparer avec une « vérité terrain », ou « *Ground Truth*, c'est-à-dire un texte de référence pour représenter, aussi précisément que possible, le contenu textuel de la source originale, afin de mesurer la précision de l'OCR obtenu en comparaison de la vérité terrain. La vérité terrain du contenu textuel d'une image consiste en l'enregistrement complet et précis de chaque caractère et mot de l'image en transcrivant manuellement et en totalité un échantillon (10 à 15 textes nous paraissent suffisant, bien que le nombre reste encore à définir). Nous détaillerons dans la dernière sous-section les différentes étapes que cela implique.

Pour mesurer la différence entre les sorties d'OCR avec la vérité terrain, nous nous appuierons sur le calcul du taux d'erreur de caractère (ou *computing error rate*, CER), car il permet précisément de mesurer la différences entre le nombre minimum d'opérations nécessaire au niveau des caractères (espaces compris) pour transformer le texte de base (ou texte de référence) en une sortie OCR. Il se base sur le calcul de la distance de Levenshtein, qui est est une métrique de distance mesurant la différence entre deux séquences de chaînes de caractères. Il s'agit du nombre minimum d'éditions (c'est-à-dire d'insertions, de suppressions ou de substitutions) d'un caractère unique nécessaire pour transformer un mot en un autre. Plus les deux séquences de texte sont différentes, plus le nombre d'éditions nécessaires est élevé, et donc plus la distance de Levenshtein est grande. Le CER se calcule comme suit :

$$CER = \frac{S + I + D}{N}$$

avec S le nombre de substitution, I le nombre d'insertion et D le nombre de suppression, sur N le nombre total de caractères dans le texte de référence. Par exemple, avec une phrase de référence comme « My name is Adrien »et une sortie « Myy neme iz Adrien », on peut compter 1 insertion, 2 substitutions et 0 suppression, soit un CER égal à $\frac{3}{17}$ soit un taux d'erreur de 17,6%. Il est a noté qu'avec cette méthode, un CER peut dépasser les 100%, comme dans l'exemple suivant, où l'on a en entrée « ABC »et en sortie « ABC12345 », ce qui nous donne un CER de $\frac{5}{3}$ soit un taux d'erreur de 166,67%, que l'on peut néanmoins normaliser avec la formule suivante :

11. https://www.primaresearch.org/www/assets/tools/Aletheia%20User%20Guide.pdf
12. https://tesseract-ocr.github.io/tessdoc/ImproveQuality.html

$$CER_{normalized} = \frac{S + I + D}{S + D + I + C}$$

où C est le nombre de caractères corrects et est toujours supérieur au numérateur, ce qui permet d'obtenir un résultat compris entre 0 et 100%. Comme l'indique cet article déjà cité [13], un bon OCR pour les archives numérisées de la presse doit avoir un CER (non normalisé) de 1-2%. Entre 2 et 10%, il s'agit d'un taux d'erreur acceptable, mais au-delà de 10%, on peut le considérer comme trop fautif. Regardons ainsi, pour quelques exemples, ce qu'il en est pour l'OCR de Gallica.

2.2.2 Pourquoi réocériser avec Tesseract ?

Nous pensons qu'il serait intéressant de tenter une ré-océrisation avec Tesseract car l'OCR de Gallica, bien qu'il indique un taux de reconnaissance de 98 mots sur 100, nous semble bien plus fautif que cela. Pour l'illustrer, nous nous sommes appuyés sur quelques extraits d'articles concernant le général Boulanger.
Pour ce faire, nous avons réalisé une ébauche de comparaison de textes océrisés avec des textes de référence en nous appuyant sur un travail réalisé dans le cadre de la TEI où nous avons, pour un petit nombre de journaux (8), extrait les articles concernant le général Boulanger (nous n'avons sélectionné que les journaux où le mot « Boulanger »apparaît au moins 5 fois) et retranscrit manuellement ceux-ci en conservant la forme originale (mots coupés, fautes de frappe, sauts à la ligne). Il ne s'agit donc pas de transcriptions de journaux entiers, mais d'une sélection d'articles des quotidiens du 8 au 17 juillet inclus (hormis celui du 16 juillet). Ils sont de longueurs et de nature diverses, la plupart situés en première page du journal. Dans nos texte de références, les lignes vides correspondent soit au passage à un article suivant, soit un changement de colonne ou de page (ce qui est de toute manière gommé dans le calcul du CER puisque l'on remplace tous les sauts de ligne par des des espaces vides). Chaque texte de référence contient en moyenne environ 14 306 caractères (pour un total de 128 760 caractères) et de 2 490 mots (comptabilisés en utilisant en remplaçant les apostrophes par des espaces puis en délimitant chaque mot par des espaces vides, pour un total de 22 406 mots), ce nombre cachant de grandes disparités (nombre de caractères minimum/maximum : 1 736/28 407 ; nombre de mots minimum/maximum : 302/4 777).

13. Holley Rose, « How good can it get ? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs », *D-Lib Magazine*, vol. 15–3/4 (2009), URL : https://www.dlib.org/dlib/march09/holley/03holley.html.

	date_extraits	ocr_output	ref_text	cer	nb_char_reftext	nb_mots_reftext
0	07-08	NE CONVERSATION POINT D'HISTOIRE ÉCLAIRCI PA) ...	UNE CONVERSATION POINT D HISTOIRE ÉCLAIRCI PAR...	3.58	3889	696
1	07-09	LE GÉNÉRAL QUITTE SA BONNE VILLE DE PARIS A l'...	VIVE BOULANGER ! LE GÉNÉRAL QUITTE SA BONNE V...	10.27	19145	3330
2	07-10	LE GENERAL BOULANGER La manifestation de vendr...	LE GENERAL BOULANGER La manifestation de vendr...	3.46	4148	650
3	07-11	Boulanger le vendredi, Pranzini" te samedi. Si...	BON VOYAGE Boulanger le vendredi, Pranzini le ...	5.44	24090	4244
4	07-12	Eh bien c'était bien la peine comme dit là cha...	L INTERPELLATION Eh bien ! c était bien la pe...	4.33	11120	1980
5	07-13	LES EMBLÈMES B.OULANGISTES ET .LA. FÊTE N.ATIO...	LE GENERAL MIRLITON LES EMBLÈMES BOULANGISTES ...	4	10615	1821
6	07-14	L'AGITATION BOULANGISTE Le 14 juillet Les mani...	L AGITATION BOULANGISTE Le 14 juillet -Les ma...	5.06	1736	302
7	07-15	COMMENT S'EST PASSÉE LA. FÊTE NATIONALE LA REV...	HIER COMMENT S EST PASSÉE LA FÊTE NATIONALE ...	5.22	28407	4777
8	07-17	EN REVENANT. Jeudi soir, au moment où les lamp...	EN REVENANT... Jeudi soir, au moment où les la...	4.95	25610	4606

FIGURE 2.6 – Capture d'écran du DataFrame du calcul du CER de nos extraits

Le taux moyen d'erreur de caractères pour l'ensemble de ces extraits s'élève à 5,15%, avec là aussi des disparités puisque l'on trouve un maximum de 10,27% et un minimum de 3,46%. Gallica estime le taux moyen de reconnaissance de mots (le taux de reconnaissance de caractères n'étant pas fourni) de ces textes à 99,49% (respectivement 99,49%, 99,4%, 99,42%, 99,55%, 99,44%, 99,55%, 99,46%, 99,55% et 99,55%). Même s'il ne s'agit pas de la même mesure, nous pouvons raisonnablement estimer que nous en sommes légèrement éloigné avec un taux de reconnaissance de caractères que nous estimons en moyenne à 94,8%. Si l'OCR de Gallica n'est pas mauvais et reste acceptable, il n'est pas non plus de très bonne qualité (notamment au niveau de la reconnaissance des caractères accentuées) et qu'une ré-océrisation via Tesseract est à tenter afin d'essayer d'obtenir de meilleur résultats.

Tesseract est un moteur logiciel OCR de reconnaissance de caractères développé à l'origine par la société Hewlett-Packard entre 1985 et 1995, qui l'a publié en open-source fin 2005, et qui est désormais sponsorisé par Google Projects (Google Tesseract). C'est un des meilleurs OCR en open-source et gratuit, avec des performances pour la reconnaissance de caractères imprimés (moins pour les caractères manuscrits) tout à fait satisfaisantes [14], et qui a également l'avantage de posséder une documentation fournie ainsi qu'une extension *Python* (pytesseract) permettant d'effectuer des manipulation plus complexes et d'automatiser la reconnaissance de caractères. De plus, Tesseract implémente des dictionnaires par défaut qui permettent de corriger en partie les sorties textes. Mais, comme nous l'avons dit, le pré-traitement des images est une étape cruciale et nécessaire qui influe considérablement sur la qualité de l'OCR. Il s'agira alors de mesurer si cela vaut la peine de ré-océriser notre corpus en comparant les sorties d'OCR d'un certain échantillon de textes avec la vérité terrain préparée au préalable. Il s'agira alors de comparer les taux moyens (ou médian, la métrique n'étant pas encore définie car elle dépendra de la disper-

14. https://research.aimultiple.com/ocr-accuracy/

sion des taux d'erreur de caractères de chaque texte) des OCR de Gallica et de Tesseract, afin de sélectionner le corpus de textes océrisés ayant le taux moyen/médian le moins élevé par rapport à la vérité terrain, corpus qui sera ensuite utilisé pour les différentes manipulations que nous faisons sur *TXM* ou *spaCy*, et que nous détaillons dans la partie suivante. Cependant, si l'OCR de Tesseract n'est que très légèrement supérieur à celui de Gallica (si, disons-le, il n'atteint pas au moins un CER d'environ 2%), nous nous contenterons très probablement de l'OCR fourni par Gallica, étant donné le caractère coûteux de la ré-océrisation.

2.2.3 Mise en place et comparaison avec la vérité de terrain

Les textes servant de références construit dans le cadre de la vérité de terrain pourront par exemple être établis à l'aide de l'outil Aletheia dont nous avons déjà parlé. Le guide du contrôle de qualité de l'OCR de Gallica [15], disponible sur leur site [16], ou encore le guide d'introduction à Aletheia [17] disponible sur le site de l'Université de Salford de Manchester [18] nous permet d'entrevoir les différentes étapes à suivre afin de pré-traiter les images pour mettre en place la vérité terrain. La binarisation des images est la première d'entre elles, ainsi que la suppression du bruit et des bordures (le logiciel propose une suppression automatique et manuelle), la création de régions (là encore, on peut automatiser celle-ci et la modifier manuellement) en leur adjoignant un attribut (texte, image etc.). Il convient ensuite de retranscrire le texte à l'intérieur de chaque région, le sens de lecture entre les différentes régions puis d'indiquer les lignes à l'intérieur de ces dernières (processus automatisés et manuels disponible). Une fois ces textes de références établis, on peut alors seulement comparer le taux d'erreur de caractères entre ces dits textes et les versions océrisées de Gallica et Tesseract, à l'aide par exemple de l'outil Text Eval, mis en place, comme Aletheia, dans le cadre du projet IMPACT.

Nous sommes néanmoins contraints de nous appuyer sur un corpus de textes océrisés imparfait et qui le sera dans tous les cas (la question est de savoir à quel point et est-ce que cela vaut la peine de ré-cériser notre corpus), et le post-traitement des textes basé sur un dictionnaire évoqué précédemment semble être la meilleure option, même avec une éventuelle meilleure océrisation par le biais de Tesseract. En attendant, nous nous contentons du corpus fourni par Gallica afin de réaliser la description et une analyse sommaire de notre corpus, que nous détaillons dans la partie suivante.

15. https://www.bnf.fr/sites/default/files/2018-11/ocr_controle_qualite.pdf
16. https://www.bnf.fr/fr/techniques-et-formats-de-conversion-en-mode-texte
17. https://www.primaresearch.org/www/assets/tools/Aletheia%20Introduction.pdf
18. https://www.primaresearch.org/tools/Aletheia/Resources

Chapitre 3

Présentation du jeu de données

Nous allons présenter dans cette partie les résultats que nous avons pu obtenir avec les différents outils que nous avons utilisés. Après avoir décrit ces derniers, nous ferons une présentation des données générales de notre corpus, puis nous nous focaliserons plus précisément sur le traitement réservé au « cas Boulanger », en nous intéressant notamment à son aspect populiste.

3.1 Deux méthodes, plusieurs outils

Pour analyser notre corpus, nous avons utilisé deux méthodes différentes, assorties chacune d'un ou plusieurs outils, que nous allons présenter : la textométrie, avec le logiciel *TXM*, et le traitement automatique des langues (TAL) avec *spaCy*, *word2vec* et *NLTK*.

3.1.1 Textométrie et TAL : différences et convergences

La textométrie est un autre nom donné à l'analyse des données textuelles (ADT), qui est une approche des sciences humaines qui envisage les textes comme des données organisées constituées en corpus. Elle est à l'intersection de nombreuses disciplines : la linguistique, l'analyse du discours, la statistique ou encore l'informatique (mais est également utilisée pour le traitement d'enquêtes socio-économiques, en marketing etc.). Pour citer les auteurs du livre « Analyse des données textuelles [1] », elle‘« combine non seulement les possibilités exploratoires et inférentielles des traitements quantitatifs des grands volumes d'archives, mais aussi la proximité au texte et la souplesse des traitements qualitatifs ». Ce champs méthodologique s'inscrit dans la continuité des travaux de statistiques lexicales, qui prennent place dès le XIXe siècle et surtout au XXe, notamment avec George Kingsley Zipf et sa célèbre loi de fréquence des mots dans un texte. Elle se développe notamment dans les années 1960 avec le déploiement de mesures et de méthodes précurseures des

1. Ludovic Lebart, Bénédicte Pincemin et Céline Poudat, *Analyse des données textuelles*, Québec, 2019, p. 2.

humanités numériques visant à décrire et caractériser le vocabulaire d'un texte ou d'un ensemble de textes, abordés selon une approche probabiliste (« sacs de mots ») où les notions d'échantillonage et d'écarts à la moyenne sont fondamentales. Cette approche a ensuite connu une nouvelle phase d'enrichissement, dans les années 1970 et 1980, avec des travaux sur l'analyse multidimensionnelle lexicale (analyses en axes principaux, méthodes de classification). Souvent confondue avec la lexicométrie, la textométrie s'en distingue pourtant par son approche du texte : là où la première s'intéresse essentiellement au lexique, concevant le texte comme un ensemble de mots, la textométrie prend en compte toutes les dimensions du texte (linéarité des unités ou séquences qui le constituent, diversité des informations linguistiques ou encore résonances fréquentielles ou cooccurrentielles des unités [2]).

Le traitement automatique des langues est quant à lui un domaine qui vise à « formaliser les descriptions linguistiques dans la perspective du développement d'une application informatique [3] », en développant des modèles simulant les mécanismes d'une langue naturelle permettant d'implémenter des applications informatiques. La visée applicative en est un des fondement, notamment avec le développement du *Web*, même si on peut dater ses origines dès les années 1950 aux Etats-Unis. Mais le TAL va être abandonné jusqu'au milieu des années 1970, au profit de l'intelligence artificielle, avec l'acquisition du traducteur *Systran* par la communauté européenne, qui stimule à nouveau la recherche (et surtout les financements). Devant les besoins croissants de traitements de documents multilingues, la traduction assistée par ordinateur est une des applications productives la plus développée, notamment avec l'essor du *Big data* et de la demande en fouille de données et de textes (indexation automatique, extraction d'informations ou encore classification des documents). Le TAL, qui traite à la fois de la langue orale et écrite, se décompose en différents niveaux d'analyse et de description que sont la morphologie, (la forme des mots, l'organisation des catégories grammaticales), la syntaxe (qui traite de la fonction des mots, étudiant la manière dont les mots se combinent pour former des phrase) et la sémantique (qui a à voir avec le sens des mots et le rapport de sens entre les mots (comme l'homonymie par exemple)) pour l'écrit.

Le TAL et la textométrie s'intéresse ainsi au même champs d'étude (entre autres, concernant le TAL), à savoir le texte et son traitement, mais diffèrent dans la manière d'aborder celui-ci : alors que le premier est centré sur la phrase, la seconde approche le texte comme un ensemble. Alors que le corpus est une source en textométrie, c'est une ressource en TAL. Elle se distinguent également par leurs objectifs : le TAL définit ses méthodes et objets relativement aux objectifs applicatifs qu'il se fixe, alors que la textométrie a d'abord

2. *Ibid.*, p. 11.
3. *Ibid.*, p. 19.

et avant tout un objectif descriptiviste et interprétatif. Cependant, des rapprochements entre les deux approches ont été réalisés, le TAL reprenant par exemple les techniques de classification et les techniques d'apprentissage de l'ADT, en sondant le corpus pour en extraire des suites de mots significatives ou des collocations. Si l'inverse est moins vérifié, la textométrie emprunte au TAL ses outils d'annotations linguistiques automatiques (étiqueteurs morphosyntaxiques, analyseurs syntaxiques). De plus, la révolution numérique et le *Web* ont permis la libération de grandes masses de données et de corpus, orientant le TAL vers les statistiques et les méthodes probabilistes, rapprochant les deux domaines autour des questions de l'usage et de la régularité. Malgré ces rapprochements, les deux méthodes poursuivent néanmoins des objectifs et applications disjointes, et les outils qui utilisent l'une ou l'autre de ces méthodes peuvent permettre une approche complémentaire d'un même corpus, ce dont nous ne nous sommes pas privés d'exploiter.

3.1.2 Les outils

3.1.3 *TXM*

Nous nous sommes appuyés sur l'outil *TXM* pour réaliser notre analyse textométrique. Initié en 2007 par le projet ANR Textométrie, lançant une nouvelle génération de recherches textométriques, le logiciel voit le jour en 2010. Il intègre notamment un moteur statistique basé sur le logiciel libre *R* (que nous n'avons pas utilisé) et un moteur de lemmatisation basé sur *TreeTager* (un outil permettant d'annoter des textes avec des informations sur les parties du discours et les lemmes). Celui-ci permet ainsi de « tokéniser »les textes (c'est-à-dire de convertir une chaîne de caractères (un texte) en une liste de symboles (*tokens* en anglais) ou lexèmes, une suite de caractères qui correspond au motif d'un symbole lexical ; le raccourci est ainsi souvent fait avec la notion de « mot ») et de les lemmatiser, c'est-à-dire appliquer aux occurrences des lexèmes sujets à flexion (en français, verbes, substantifs ou encore adjectifs) un codage renvoyant à leur entrée lexicale commune que l'on appelle lemme (le lemme du lexème « est »est « être »). *TreeTager* attribue également à chaque *token* une étiquette morphosyntaxique, ou autrement dit, la classe grammaticale, ce qui n'est pas forcément une chose aisée (« été »peut par exemple renvoyer à la troisième personne du singulier du verbe « être »ou au nom commun désignant la saison) et est souvent imparfaite, nous aurons l'occasion de le constater. Ce moteur permet construire et à analyser tout type de corpus textuel numérique, éventuellement étiqueté et structuré en XML, et, à l'instar de la méthode qui le sous-tend, permet d'explorer un corpus quantitativement (études des cooccurrences, classification etc.) et qualitativement, en permettant un retour au texte grâce au mode « édition »ou encore le concordancier. Par sa disposition en tableau, la concordance permet en effet l'observation systématique et complète des contextes locaux (dans une fenêtre prédéfinie de mots de part et d'autre du mot pivot), permettant de repérer des régularités et des récurrences.

C'est également une forme synthétique et dense d'affichage des contextes d'usage de toutes les occurrences d'un mot dans le corpus, effaçant ainsi les distances entre les occurrences. Nous en montrons un exemple concret dans la section suivante.

Nous avons implémenté dans le logiciel deux versions du corpus : la version brute, sans nettoyage du texte ni pré-traitement (suppression des chiffres et des mot-outils (mots dont la fonction est plutôt grammaticale que sémantique, comme les déterminants, les pronoms et les prépositions)), puis la version nettoyée et pré-traitée. L'objectif était ainsi de comparer les deux versions, et notamment de l'impact du corpus pré-traité sur la l'étiquetage morphosyntaxique des *tokens*, dont nous discutons dans la section suivante. Nous avons également importé un fichier de métadonnées dans lequel nous avons renseigner l'identifiant des textes, basé sur le nom des fichiers, la date de publication, le nom du journal, le nombre de pages par texte ainsi que le format du journal (longueur et largeur en millimètres). Cette dernière information a notamment pu être récupérée sur l'API de la BNF [4] qui met à disposition les métadonnées quantitatives de la presse ancienne. Ces métadonnées renferment d'autres informations précieuses comme le nombre d'articles, le nombre de mots, le nombre de blocs de texte ou encore le nombre d'illustrations, mais étant donné que nous nous sommes bornés aux deux premières page de chaque texte, ces informations étaient de fait inutilisables pour notre corpus. Le choix des métadonnées n'a pas été fait au hasard et s'inscrit dans la perspective d'une étude d'un plus large corpus de journaux (d'où l'intérêt de renseigner le nom du journal et le nombre de pages) ; quant au format, il n'est pas sans importance, et constitue un cadre rigide dans lequel un journal est produit, ce qui a des conséquences que nous n'analyserons pas ici, mais qui pourront à l'avenir être très utiles.

3.1.3.1 Les outils du TAL

Nous utilisons trois outils dans le cadre du TAL : *spaCy*, *word2vec* et *NLTK*.

SpaCy est une bibliothèque logicielle *Python* qui permet de réaliser des recherches approfondies sur un texte, mais également de prendre en charge les processus d'apprentissage profond qui permettent de connecter des modèles statistiques formés par des bibliothèques d'apprentissage automatique populaires comme *TensorFlow*. A l'instar de *TXM*, il permet la tokénisation, la lemmatisation et l'attribution d'étiquettes morphosyntaxiques aux mots du corpus, et pour le français plus spécifiquement, le réalise par le biais de Le*fff* (Lexique des Formes Fléchies du Français). Cet outil nous permet ainsi de réaliser des analyses précises du texte, notamment grâce à son outil de reconnaissance d'entités nommées poussé (nous y reviendrons) et, en ce qui nous concerne, de visualiser à la fois les composantes

4. https://api.bnf.fr/fr/metadonnees-quantitatives-de-la-presse-ancienne-xixe-xxe-siecles

générales de notre corpus (le nombre de mots par fichier par exemple), mais également des composantes plus précises (les occurrences du terme « Boulanger »par exemple.

Word2vec est une technique du TAL éditée en 2013. L'algorithme *word2vec* utilise un modèle de réseau de neurones pour apprendre les associations de mots à partir d'un large corpus de textes. Une fois entraîné, un tel modèle peut détecter des mots synonymes ou suggérer des mots supplémentaires pour une phrase partielle. Comme son nom l'indique, *word2vec* représente chaque mot distinct par une liste particulière de chiffres appelée vecteur, permettant ainsi d'être interprétés par un algorithme. Les vecteurs sont choisis avec soin de manière à ce qu'une simple fonction mathématique (la similarité en cosinus entre les vecteurs) indique le niveau de similarité sémantique entre les mots représentés par ces vecteurs. La vectorisation de mots (ou *words embedding* en anglais) est ainsi au centre de cette technique : les mots apparaissant dans des contextes similaires possèdent des vecteurs correspondants qui sont relativement proches (comme « roi »et « reine »par exemple). Cette technique est basée sur l'hypothèse dite « de Harris »qui veut que les mots apparaissant dans des contextes similaires ont des significations apparentées et de maximiser la probabilité d'observer un mot cible en relation avec les autres mots qui l'entourent (mot de contexte).

La librairie *NLTK* (pour *Natural Language Toolkit*) est une bibliothèque logicielle en *Python* destiné à soutenir la recherche et l'enseignement en TAL ou dans des domaines étroitement liés, y compris la linguistique empirique, les sciences cognitives, l'intelligence artificielle, la recherche d'information et l'apprentissage automatique. Comme *spaCy*, elle permet de tokéniser, lemmatiser et d'attribuer des étiquettes morphosyntaxiques ou encore la reconnaissance d'entités nommées, mais également l'analyse de sentiment ou la segmentation en sujets thématiques (ou *topic*). Il permet encore d'autres manipulation comme celle du calcul des ngrammes (des suites de n caractères ou mots), permettant dans notre cas de visualiser les suites de mots les plus fréquentes (même si nous nous sommes essentiellement concentrés sur les bigrammes, les suites de deux *tokens*).

3.2 Description générale du jeu de données

3.2.1 Test sur *TXM* avec le corpus brut et le corpus nettoyé

Comme nous l'avons énoncé précédemment, le but de la démarche était de tester la lemmatisation de *TXM*. En effet, en nous appuyant sur un corpus pré-traité, sans les mot-outils, l'étiquetage morphosyntaxique, qui dépend aussi du contexte du *token*, peut s'en trouver affecté à la perte. Nous n'avons pas réaliser une étude systématique des différences d'étiquetage entre le corpus brut et le corpus nettoyé et pré-traité, aussi nous

sommes nous focaliser sur l'exemple précis de l'étiquetage du *token* « Boulanger ». *TXM* propose en effet une reconnaissance des noms propres (étiquette « *NAM* ») et nous désirons ainsi voir si le corpus brut permet de significativement mieux étiqueter notre objet.

En nous intéressant au lemme « Boulanger »(en ignorant le fait que les caractères soient ou non en majuscule avec la commande %c), afin de prendre en considération ses différentes déclinaisons possibles, nous avons ainsi pu constater que l'étiquetage était légèrement meilleur avec le corpus brut : 1107 occurrences du terme sur tout le corpus contre 1055 avec le corpus pré-traité, soit une perte de 4,70% pour ce dernier. Nous avons considéré qu'en dessous du seuil de 5%, il ne s'agissait pas d'une perte significative et avons ensuite effectué toutes nos manipulation sur le corpus pré-traité. De plus, si globalement, l'étiquetage est meilleur avec le corpus brut, il y a néanmoins également des pertes au niveau du corpus brut par rapport au corpus pré-traité, que l'usage du concordancier a pu mettre en évidence. Nous pouvons en effet constaté que le terme « BOULANGER »du 19 juillet est bien indexé en tant que *NAM* dans le corpus pré-traité alors qu'il ne l'est pas dans le corpus brut (voir la figure 3.2).

Requête [frlemma="Boulanger" & frpos="NAM"]

text_id	Contexte gauche	Pivot	Contexte droit
lematin_18870717	Oh t entends bien Laur collaborateur Lalou ami Laur général	Boulanger	écrit point responsable publicité ami compagnon voyag...
lematin_18870717	dss Débats lettre cire puljlic c assentiment écrite Do chocs	Boulanger	autorise expressément tacitement publication agit acte i...
lematin_18870717	tacitement publication agit acte indiscipline flagrant Laur ab...	Boulanger	commis acto légèreté impardonnable prenant journa lisl...
lematin_18870717	portée règlements étala discipline militaires Siècle met dout...	Boulanger	prêt fairo devoir désire France respectée G ost cas offi...
lematin_18870717	dise do rolever telo Kilo relevée songe faire baisser général	Boulanger	enferme devoirs commandement corps armée trouve ra...
lematin_18870717	illusion combinaison ministérielle Justice contente do publier...	Boulanger	Voltaire publie dernière heure information snivante Héri...
lematin_18870717	publie dernière heure information snivante Hérisse vient rec...	Boulanger	résulte lettre adressée Francis Laur caractère intime gé...
lematin_18870718	pourtant criaient Vive Gambetta nez MacMahon difficiles mu...	Boulanger	nez respectable Grévy disant choses permettant rappro...
lematin_18870718	MATIR ¡ LETTRE lit XIX Siècle lu hier lettre général	Boulanger	écrite Laïïr celni livrée publicité publication été vivemen...
lematin_18870718	aucun subordonnes écrire àses amis règlements militaires é...	Boulanger	ministre guerre eau cousidération faire valoir Chambre ...
lematin_18870718	parlementaire écartée demander raison destinataire lettre p...	Boulanger	insulte croire recours moyen tourner règlements somm...
lematin_18870718	crier défendait crier contre déclarait publiquement mot ordre...	Boulanger	ligueurs crié Ferry répondre défendu crier contre memb...
lematin_18870718	crier contre membres gouvernement Jules Ferryn membre g...	Boulanger	gouvernement directeur vrai chef soa trompaientJules ...
lematin_18870719	clercs fins JULES DELAFOSSE MATIN publiera demain arti...	BOULANGER	télégramme commandant S corps Laur commandant e ...
lematin_18870719	télégraphie avez reproduire France lettre ignoriez doute droi...	Boulanger	Francis Laur pousse farce faire télégramme général Pé...
lematin_18870719	création été commise ministère sujet dépêche privée adress...	Boulanger	Laisant député paru utile faire pénétrer lecteurs fameux...
lematin_18870719	partis agitaient questions personnes amplement raison répu...	Boulanger	Philippe VII leader blanc chaut combats gouvernement f...
lematin_18870723	bien inspirés hier acceptant bénéfice inventaire racontars sy...	Boulanger	Laur débite jours tréteaux obligeamment mis disposition...
lematin_18870723	demain députés criminels nierait énergie moyen aurions dé...	Boulanger	réduit silence porter garant donnant noms courons dén...
lematin_18870723	donnés X X invite général Ferron entrer scène faire général	Boulanger	pantalonnade édifiante franchement duréCHAMBRE B...
lematin_18870723	élucubrations X X FIL SPÉCIAL LONDRESjuillet Morning Po...	Boulanger	chargé commandement corps armée pris aucune part é...

FIGURE 3.1 – Concordancier du corpus pré-traité autour du terme « Boulanger »

Requête [frlemma="Boulanger" & frpos="NAM"]

text_id	Contexte gauche	Pivot	Contexte droit
lematin_18870718	a lu, hier, la lettre que le général	Boulanger	a écrite à M. Laïïr. et que celni
lematin_18870718	règlements militaires n'ont pas été violés par le général	Boulanger	, et le ministre de la guerre n'aurait eau
lematin_18870718	la publier. Nous ne voulons pas faire au général	Boulanger	l'insulte de croire qu'il a eu recours à
lematin_18870718	son mot d'ordre qu'en même temps que Vivo	Boulanger	I les ligueurs avaient crié : A bas Ferry !
lematin_18870718	a eu des moments où les am's de M.	Boulanger	ont dit que le gouvernement avait pour directeur et pour
lematin_18870719	quoi que ce soit dans les journaux. 9 Général	Boulanger	. » M. Francis Laur ne pousse pas.
lematin_18870719	une dépêche privée adressée par M. le géné- ral	Boulanger	à M. Laisant, député, il nous a
lematin_18870719	même degré. Que ce soit Urgéle ou Morgan,	Boulanger	ou Philippe VII, tel leader de tel blanc ou
lematin_18870723	les racontars que le syndicat X. X.,	Boulanger	, Laur et Ce débite depuis deux jours sur les
lematin_18870723	de démontrer la réalité d'un fait dont le général	Boulanger	, réduit au silence, peut seul se porter garant
lematin_18870723	à entrer en scène et à faire parler le général	Boulanger	. La pantalonnade est édifiante, mais, franchement,
lematin_18870723	Le Morning Post ne peut pas admettre que le général	Boulanger	, qui est chargé du commandement d'un corps d'
lematin_18870723	de coup d'Etat qui auraient été faites au général	Boulanger	, ministre de la guerre. M. le général
lematin_18870723	, ministre de la guerre. M. le général	Boulanger	a de terribles amis, plus terri. bles,
lematin_18870723	semble avoir été écrite après une conversation avec le gén...	Boulanger	, nous ne croyons pas aux faits qu'elle révèle
lematin_18870723	il n'est pas admissible que M. le général	Boulanger	ait gardé pour lui ces propositions et se soit contenté
lematin_18870724	convaincus qu'on a abusé de confidences intimes du général	Boulanger	, qu'on a dénaturé ses paroles et qu'il
lematin_18870725	plus mal, et les journaux de Paris qui oublient	Boulanger	pour lui faire de la réclame, n'arriveront pas
lematin_18870725	. 196 Boudreau, radie. 147 Capron. 17	Boulanger	. Lesueur, conservateur. 10 Divers 53 Jjallottage.
lematin_18870725	jeter aux radicaux un cri d'aJarmc. Le général	Boulanger	est de nouveau mis en ayant par eux, et
	les 91 généraux sont simplement 'allés dire au général	Boulanger	qu'ils so battraient en cts de ftuoii ",

FIGURE 3.2 – Concordancier du corpus brut autour du terme « Boulanger »

3.2.2 Données générales

Notre corpus est composé de 1 908 793 *tokens* dans le corpus nettoyé et pré-traité (5 037 540 pour le corpus brut) dont on peut voir la distribution par fichier ci-dessous :

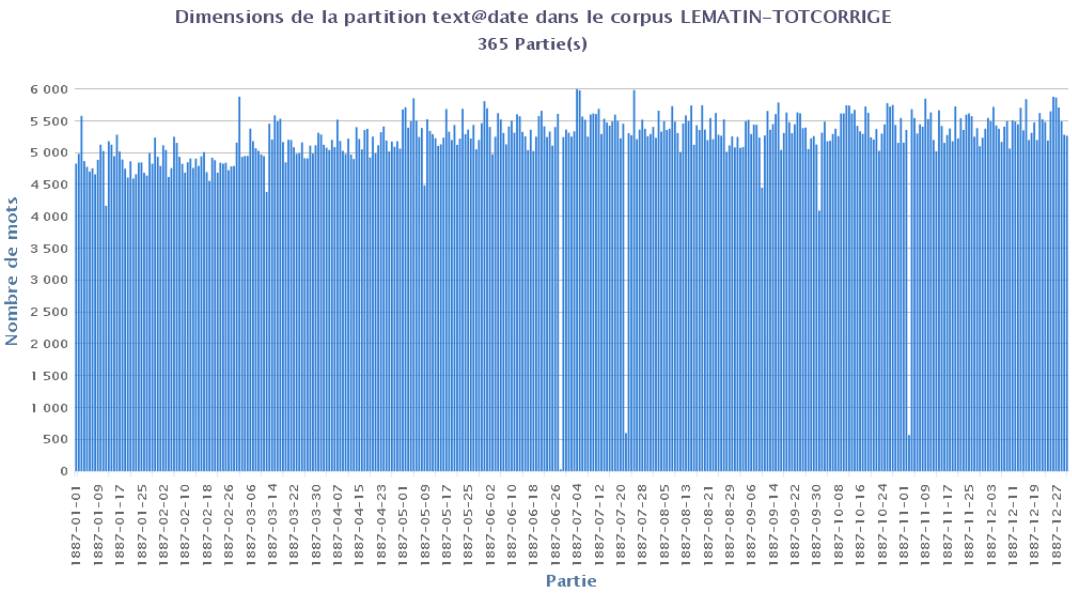


FIGURE 3.3 – Nombre de mots par fichier. Graphique exporté de TXM

On peut s'apercevoir qu'il y a quelques données scélérates dont nous n'avions pas connaissance, à savoir les fichiers des dates du 06-28, du 07-22 et du 11-03, et que nous corrigerons ainsi à l'avenir. Chaque fichier est composé en moyenne d'environ 5 230 *tokens* (5229,57 précisément). Nous avons comparé la tokénisation de *TXM* avec celle de *spaCy*, et on peut voir que ce dernier en dénombre un peu plus (1 944 128 *tokens*, soit 1,85% de *tokens* supplémentaires), mais cela est très probablement dû au fait que nous avons légèrement modifié notre fichier de nettoyage de textes après avoir implémenté et réalisé différentes mesures avec le corpus implémenté dans *TXM*. C'est d'ailleurs par le biais du retour au texte et du concordancier que nous avons pu visualiser ces erreurs de codage : en effet, nous avons retiré manuellement les « M. »(« Monsieur ») et « MM. »(« Messieurs ») ; avec notre première version de pré-traitement, cela fusionnait deux *tokens* en un seul, ce qui à notre sens peut expliquer au moins en partie ce différentiel. La moyenne du nombre moyen de *tokens* par fichier est ainsi de 5 326 (5 326,38 précisément). En enlevant les valeurs scélérates suscitées, on atteint un nombre moyen de 5 367 (5 366.82 précisément) avec *spaCy* (mais nous n'avons pas pu effectuer ce calcul avec les données de *TXM* car, pour une raison indéterminée, l'export de ce fichier (et uniquement celui-ci) produisait un fichier vide).

Nous avons ainsi pu extraire un nuages de lemmes à partir de *TXM*. Nous aurions aimé pouvoir le comparer avec la lemmatisation de *spaCy*, mais nous ne sommes pas parvenu à réaliser ce décompte malgré nos multiples tentatives. L'idée aurait été de pouvoir com-

parer la lemmatisation entre les deux outils, mais nous n'y sommes pas parvenu. Il faut tout d'abord ici préciser une erreur de notre part lors du pré-traitement du texte, celle de n'avoir pas retiré manuellement le *token* « do », qui est la version mal océrisée de la préposition « de », et qui est qui plus est une erreur récurrente (le déterminant « le »est de la même façon mal reconnu, mais de manière beaucoup moins systématique). Hormis les verbes « être »et « faire », dont la forte fréquence est peu étonnante, on peut constater avec ce nuage de mot la forte présence du lemme « général », qui ne renvoie bien sûr pas uniquement à Boulanger (mais peut aussi référer à son successeur au ministère de la guerre, le général Ferron), avec 6 771 occurrences, et qui est ainsi devant des mots qu'on imaginerait plus communs comme « bien »ou « grand ». Mais cela ne nous permet en rien d'inférer quoique ce soit à propos de Boulanger lui-même, et c'est pourquoi nous revenons plus précisément sur son cas ci-après.

FIGURE 3.4 – Nuage de lemmes réalisé avec *Python* après extraction des données de *TXM*

3.3 Boulanger : une présence importante mais contrastée

3.3.1 Le général sous le feu des projecteurs

SpaCy comme *TXM* possèdent tous deux un étiquetage des noms propres (plus précisément les entités nommées avec *spaCy*, mais le premier a l'avantage par rapport au second de pouvoir différencier les différents types d'entités : lieux, personnes etc. Ainsi, afin de se concentrer sur la comparaison entre les personnes, nous avons ici préféré utiliser *spaCy*, malgré ses défauts, nous y reviendrons. On peut voir à travers le graphique ci-

dessous la prépondérance du général qui apparaît en tout à 784 reprise si l'on tient compte de l'entité « général Boulanger »qui apparaît 444 fois et « Boulanger »qui apparaît 340 fois. On voit tout de suite une des limites de la reconnaissance d'entités de *spaCy* qui opère une distinction entre les deux entités, mais les résultats restent néanmoins exploitables. En tenant compte de ces deux entités, on voit que le général apparaît plus que Bismarck ou encore les présidents du conseil Goblet et Rouvier. Ce qui peut s'expliquer également par le fait que le premier est démis de ses fonctions en mai, remplacé par le second. Mais Boulanger est évincé de son ministère (qui est qui plus est un ministère de seconde zone) en même temps que Goblet, et il continue pourtant de faire parler de lui. Mais contrairement à ce dernier, le général est un séducteur, qui veut plaire et être populaire [5], et pense pouvoir imposer son retour à la classe politique grâce au soutien de l'opinion publique [6].

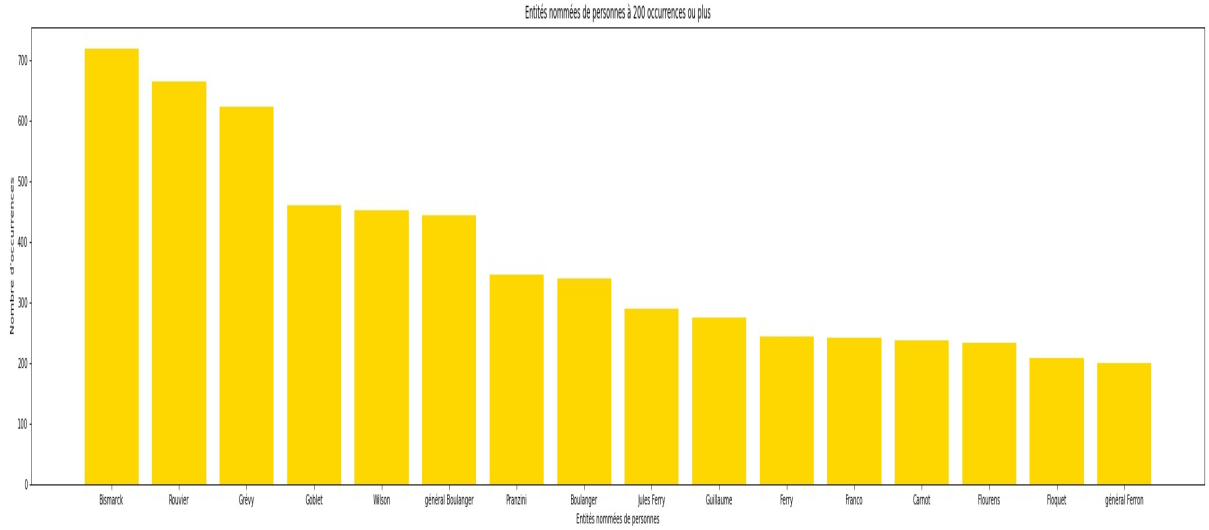


FIGURE 3.5 – Fréquence des entités nommées de personnes de 200 occurrences ou plus

3.3.2 Des apparitions disparates

Afin de visualiser la présence du général dans le temps, nous nous sommes appuyés à la fois sur les données fournies par l'étiquetage des noms propres de *TXM* et les *tokens* « Boulanger »de *spaCy*. On peut observer le chiffre relativement proche entre les deux nombres : 1 055 pour *TXM*, 1 120 pour *spaCy*, soit 30% de perte au niveau de la reconnaissance d'entités nommées pour ce dernier (qui en comptabilisait 784). Bien qu'ils soient de natures différentes (étiquetage morphosyntaxique pour l'un, reconnaissance des entités pour l'autre), on peut néanmoins constater la plus grande pertinence de l'étiquetage des

5. B. Joly, *Aux Origines du populisme. Histoire du boulangisme (1886-1890)*..., p. 124.
6. *Ibid.*, p. 218.

noms propres de *TXM* par rapport à la reconnaissance d'entités nommées de *spaCy*.

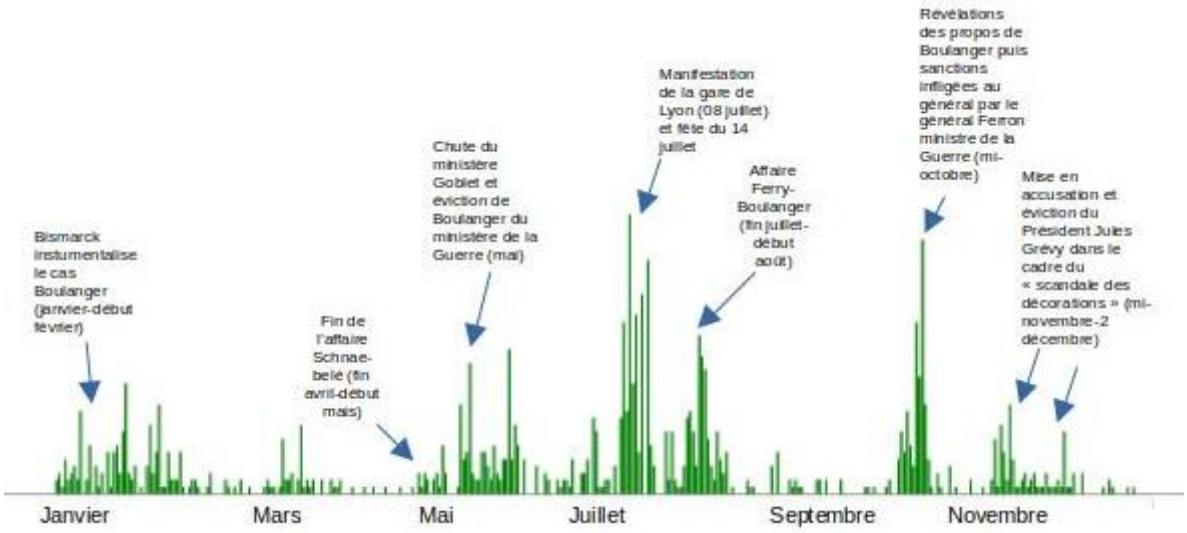


FIGURE 3.6 – Nombre d'occurrences du *token* « Boulanger »par quotidien

On distingue clairement cinq périodes où le général suscite une attention particulière : de janvier à début février ; au mois de mai (et début juin) ; en juillet jusqu'à la mi- août ; en octobre pendant une dizaine de jours ; de la mi-novembre à début décembre.
La première période coïncide avec les débuts du nouveau ministère qui a pour Président du Conseil Goblet, tout juste constitué en décembre 1886 : Boulanger est déjà perçue comme une menace et est proche d'être écarté du nouveau gouvernement, mais sa popularité grandissante le sauve, pour un temps du moins. Car le général inspire de plus en plus la crainte, renforcée par les assertions de Bismarck à son égard, qui instrumentalise le cas Boulanger et sa rhétorique pro-revancharde (qui tient plus de la posture que d'une réelle volonté d'attaquer l'Allemagne) pour menacer la France en cas de début de velléités belliqueuses. Cependant, il ne faut pas s'y tromper : d'une part, la revanche est un concept beaucoup moins transparent qu'il n'y paraît : il ne s'agit pas d'une volonté de reconquérir de manière offensive les provinces perdues, mais plus d'affirmer un sentiment de défiance, autant que de crainte, envers le voisin germanique. Le pacifisme reste de mise, même si l'on oublie pas les provinces perdues, et s'il se mêle à un patriotisme ardent que le général saura exploiter, la revanche demeure un concept purement défensif [7]. D'autre part, le général Boulanger, s'il est partagé entre l'envie et la prudence, n'a jamais donné de signes tangibles du déclenchement d'un nouvelle guerre avec l'Allemagne, quoi qu'aient pu en dire ses détracteurs ou Bismarck lui-même [8]. De la même manière, les accusations d'un possible coup-d'Etat de la part du général, qui alimente d'ailleurs les pages des quotidiens, notamment celui du *Matin*, relève plus du fantasme que d'une menace réelle. Le général, plus préoccuppé par des considérations tactiques (crainte de l'échec, retombées

7. *Ibid.*, p. 30-31.
8. *Ibid.*, p. 292-294.

dangereuses) que par les lois républicaines, s'est toujours refusé à toute tentative, alors qu'il en a pourtant eu trois fois la possibilité [9] , et ce dès décembre 1887, lors du scandale des décorations affectant Jules Grévy et son gendre, Daniel Wilson.

La seconde période correspond à la chute du ministère Goblet et le départ du général du gouvernement. L'affaire Schnaebelé (du nom d'un espion français arrêté à la frontière allemande) en avril-mai, façonne la légende d'un Boulanger intransigeant face à l'Allemagne, ce qui suscite une crainte grandissante à son égard, et entraîne son éviction en mai après le renversement du ministère Goblet. Les opportunistes, rompant leur fragile alliance avec les radicaux, agissent cette fois pour bouter le général hors du champs politique, faisant fi de sa popularité. Ses sympathisants, toujours plus nombreux, n'acceptent pas cette manoeuvre, et le font notamment savoir lors de la manifestation de la gare de Lyon le 8 juillet, au moment du départ du général vers son nouveau lieu d'affectation à Clermont-Ferrand, ayant pour charge le 13e corps d'armé et qui correspond à la troisième période de forte présence du général, relative à plusieurs évènements : la manifestation de la gare de Lyon le 8 juillet, qui ne manque pas d'attiser encore un peu plus les craintes de la presse devant la popularité du général ; ses démarches en marge de la fête du 14 juillet afin de maintenir l'attention sur lui ; et l'imbroglio avec son ennemi Jules Ferry, qu'il a provoqué en duel après que ce dernier ait « atteint son honneur » lors d'un discours à la Chambre (duel qui n'aura pas lieu et dont Le Matin relate toutes les péripéties). En octobre, le journal relate les évènements liés aux sanctions infligées par le nouveau ministre de la guerre à son prédécesseur, en conséquence des propos que celui-ci a tenu dans des interviews (non-autorisées) dans les journaux du *Matin*, de *Gil Blas*, de la *Nation* et du *Soir*. *Le Matin* profite ainsi de n'importe quelle occasion pour parler du général, et en fait à l'occasion plusieurs articles dans le même journal (quatre dans l'édition du 11 juillet), en fait l'objet de l'éditorial politique du journal, comme dans l'édition du 12 juillet, ou encore l'article de la colonne centrale de la première page (comme dans l'édition du 13 juillet). Enfin, entre fin novembre et début décembre, le président de la République Jules Grévy est poussé à la démission à cause du « scandale des décorations »impliquant son gendre Daniel Wilson, travaillant alors à l'Elysée. Ce dernier est en effet accusé de vendre des décorations et médailles décernées par l'Etat au prix fort et monnaye aussi son intervention pour obtenir des marchés publics ou la grâce présidentielle. Le nom de Boulanger est alors propagé pour remplacer Grévy à la tête de l'exécutif voire pour renverser le régime.

Cependant, le graphique ci-dessous semble à première vue tempérer cette posture : on

9. J. Garrigues, « Le Général Boulanger et le fantasme du coup d'État », *Parlement[s], Revue d'histoire politique*–12 (2009/2), URL : https://www.cairn.info/revue-parlements1-2009-2-page-43.html, p. 43-48.

constate que le général (du moins les entités nommées que nous avons choisies) n'apparaissent aucune fois dans 136 quotidiens. Si l'on compte les fois où le général est nommé moins de cinq fois, on peut voir que ce total monte à 292, soit 80% du total. Finalement, le général ne sature pas non plus totalement l'espace médiatique, ou seulement à quelques occasion bien précise, pendant lesquelles il est particulièrement ciblé (il cherche à l'être, d'ailleurs), voire est l'objet principal de l'édition (le 11 juillet avec 41 occurrences ou encore le 16 octobre avec 37 pour les deux valeurs extrêmes).

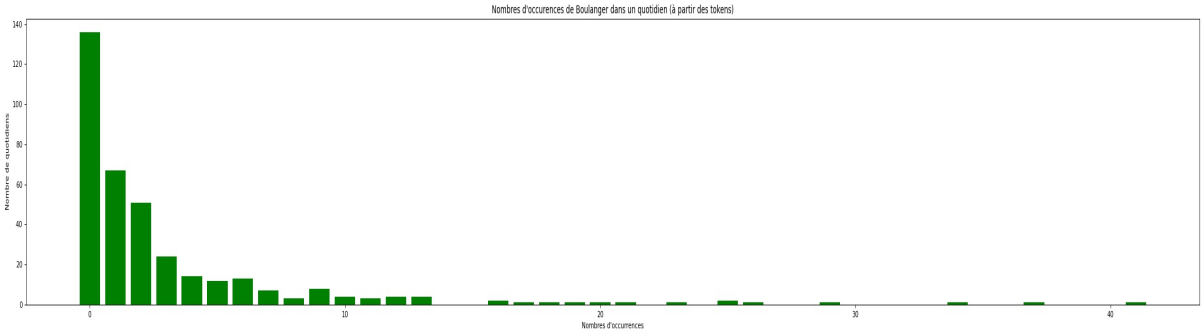


FIGURE 3.7 – Nombre de quotidiens contenant n occurrences du *token* « Boulanger »

Nous avons comparé ces chiffres à quatre autres personnages politiques, afin de confronter ces chiffres à d'autres entités nommées, en nous appuyant sur les exports de *TXM* que nous avons pu récupérer : Otto von Bismarck, Georges Clémenceau (chef des républicains radicaux) Jules Grévy (président de la République) et René Goblet (président du conseil de janvier à mai 1887).

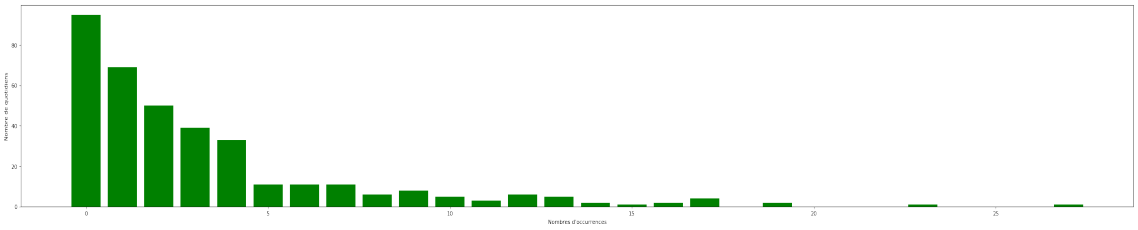


FIGURE 3.8 – Nombre de quotidiens contenant n occurrences du *token* « Bismarck »

On peut observer dans la figure ci-dessus que Bismarck est nommé dans beaucoup plus de quotidiens, avec 95 d'entre eux dans lesquels il n'est pas cité, mais si l'on s'intéresse au nombre de fois où il apparaît moins de 5 fois, on obtient un chiffre proche de celui du général (286 quotidiens). De plus, on peut constater que les valeurs extrêmes sont beaucoup moins élevées (max = 27) et globalement, le nom de Bismarck est cité entre 5 et 10 fois par quotidiens (cela représente 65,8% des quotidiens pour lesquels il est cité 5 fois ou plus). Ainsi, Bismarck bénéficie comme Boulanger d'une exposition continue dans le journal, plus souvent évoqué (dans le sens « cité au moins une fois ») que ce dernier du fait de son importance géopolitique forte. Mais, à part quelques quotidiens qui se sont

focalisés sur lui (probablement durant l'affaire Schnaebelé), il semble plus s'agir d'une exposition continue mais moins ciblée que pour le général.

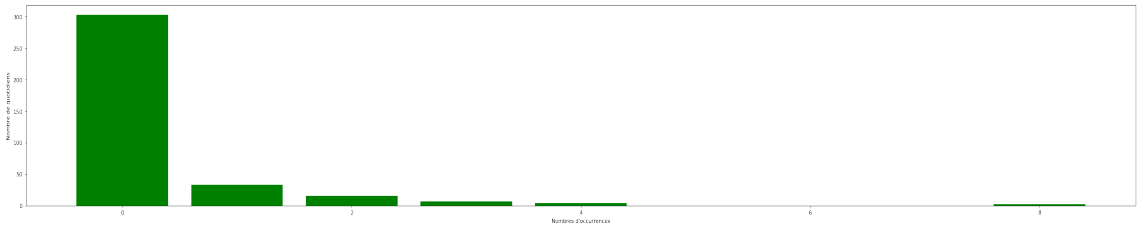


FIGURE 3.9 – Nombre de quotidiens contenant n occurrences du *token* « Clémenceau »

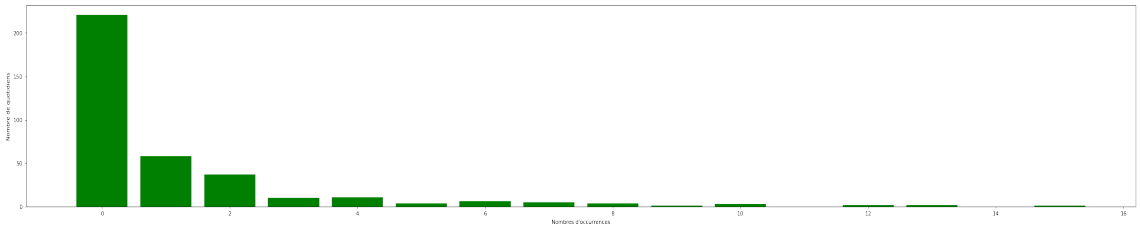


FIGURE 3.10 – Nombre de quotidiens contenant n occurrences du *token* « Goblet »

Georges Clémenceau, chef de file des républicains radicaux, pourtant souvent mis en avant lors de débats parlementaires ou du fait de ses nombreux démêlés avec les opportunistes, n'apparaît ici que dans un nombre extrêmement réduit de cas. Dans 303 quotidiens, il n'est même pas cité, et dans seulement 2 d'entre eux, il est cité plus de 5 fois (8 fois). Son président du conseil René Goblet est dans une situation relativement similaire puisqu'il est absent du quotidien 221 fois (337 fois 5 fois ou moins, soit 92,33% du total), et n'obtient que 8 fois 10 occurrences ou plus (probablement lors du renversement de son ministère en mai). Il n'apparaît quasiment plus après le mois de mai, hormis en fin novembre-début décembre lors du « scandale des décorations ». On voit ainsi que ces deux personnages très importants (l'un allié du général, l'autre son supérieur hiérarchique jusqu'en mai) génère beaucoup moins d'attention (et de manière beaucoup moins importante lorsque c'est le cas) par rapport à Boulanger.

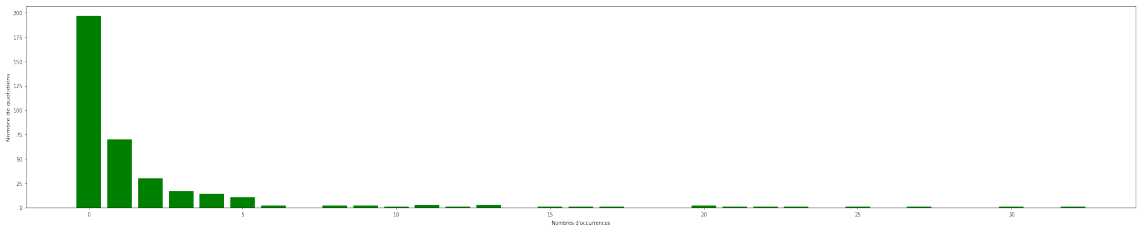


FIGURE 3.11 – Nombre de quotidiens contenant n occurrences du *token* « Grévy »

Enfin, le président de la République Jules Grévy lui-même n'apparaît pas dans la plupart des quotidiens (197 d'entre eux à 0 occurrences, 328 fois à moins de 5 occurrences soit

89,86% du total). Au-dessus de 5 occurrences, on observe des effectifs de un ou deux quotidiens par nombre d'occurrences, avec deux valeurs extrêmes à 30 et 32 occurrences (lors du « scandale des décorations »). Hormis cet épisode très précis où le président monopolise l'attention, on peut ainsi observer que sa présence est très discontinue.

Ces différents exemples nous permettent de tempérer notre propos précédent : certes, le général apparaît de manière disparate, mais c'est le cas de tous les exemples cités et, hormis Bismarck, ils sont tous moins souvent cités et avec des fréquences extrêmes nettement moindres (y compris Bismarck) que pour Boulanger. La distribution d'apparitions de Boulanger semble même anormale en comparaison de ceux-ci dans le sens où celle-ci suit une courbe beaucoup plus étalée, et avec un nombre moins élevé de fois où il apparaît moins de 5 fois. Ainsi, cette petite rétrospective tend plutôt à confirmer notre intuition générale : Boulanger est un thème majeur du quotidien *Le Matin*.

3.4 Une dynamique populiste difficile à mesurer

Pour tenter de mesurer la dynamique populiste du général, nous avons tenté plusieurs manipulation avec *word2vec* et *NLTK*, mais aussi avec les cooccurrences de *TXM*, afin de tenter d'identifier des associations de termes pouvant nous permettre (ou non) d'aller dans ce sens.

3.5 *NLTK* : recherche des collocations les plus fréquentes

Nos quelques tests sur les ngrammes, et plus particulièrement les bigrammes, n'a pas donné de résultats très intéressant et c'est la raison pour laquelle nous ne nous attarderons pas dessus. Nous pouvons notamment faire ressortir de figure ci-dessous la prépondérance du général Boulanger dans le corpus, qui détient les deux premiers bigrammes que l'on pourrait dire significatifs ((mnistère, guerre), (général, Boulanger)), même si ce dernier ne le désigne pas nécessairement, les trois premiers ((SERVICE, SPÉCIAL), (FIL, SPÉCIAL) et (SPÉCIAL, LONDRES)) étant plutôt liés à la mise en page du journal qui insiste, nous l'avons déjà noté, sur sa rapidité et sa connexion au flux d'information mondial comme argument de vente. Nous avons par ailleurs essayé de calculer les bigrammes autour du général Boulanger (voir le notebook), mais avons finalement estimé que ces résultats n'étaient pas très exploitables, faute de normalisation des valeurs sur une échelle de 0 à 1 (ou de 0 à 100).

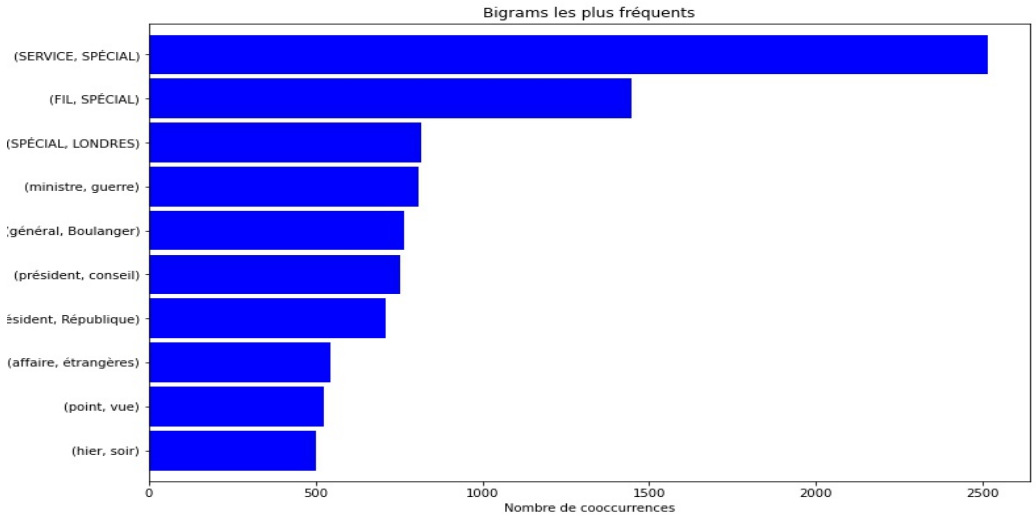


FIGURE 3.12 – Bigrammes les plus fréquents dans le corpus

3.5.1 Les cooccurrences avec *TXM*

L'outil *TXM* permet de travailler plus précisément et facilement sur le texte, notamment avec sa commande « Cooccurrences ». Elle permet, par le biais d'une requête CQL, de s'intéresser au voisinage d'un mot pivot, ou autrement dit, obtenir des mesures synthétiques (cofréquence, indice, distance moyenne) mais aussi qualitatives avec le retour au texte. La requête s'effectue autour d'un pivot (mot ou groupe de mots) et dans un contexte prédéfinie, généralement de l'ordre d'une dizaine de mot de part et d'autres du pivot. Mais étant donné que dans notre cas, nous avons étudions un corpus duquel nous avons retiré les mots-outils, nous estimons qu'un contexte de 5 de part et d'autre du pivot est suffisant. Nous avons choisi de représenter les cooccurrents de certains termes (« Boulanger », en contexte 10 cependant, et « populaire »), groupes de mots (« coup Etat ») et groupes de termes (« dictat. »pour « dictature »ou « dictateur »). Nous avons choisi de représenter, parmi les mesures énoncées précédemment,, deux d'entre elles, la cofréquence et l'indice. La première correspond au nombre de fois où les les deux termes, le pivot et le cooccurrent, apparaissent dans la fenêtre de contexte prédéfinie. Par exemple, on peut voir avec la figure que, dans une fenêtre de cinq de part et d'autre du pivot « populaire », le cooccurrent « Boulanger »n'est cofréquent avec le mot pivot que six fois. La deuxième mesure est un indice de spécificité permettant de rechercher les mots dont la fréquence autour du pivot est anormalement élevée, « compte tenu de leur fréquence générale, dans le corpus et de la taille de la partie formée par les contextes, représentants les « occasions »de cooccurrences [10] ». Afin de mieux les repérer, le *token* « Boulanger »est représenté en rouge et le *token* « général »en vert, même si celui-ci ne désigne pas uniquement Boulanger, mais

10. L. Lebart, B. Pincemin et C. Poudat, *Analyse des données textuelles*..., p. 120.

aussi le général Ferron.

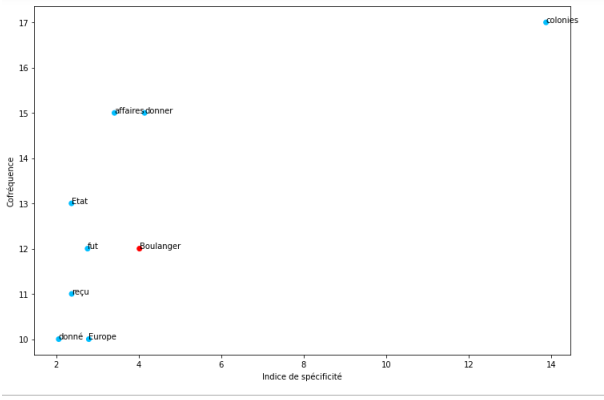


FIGURE 3.13 – Cooccurrents de « coup Etat »

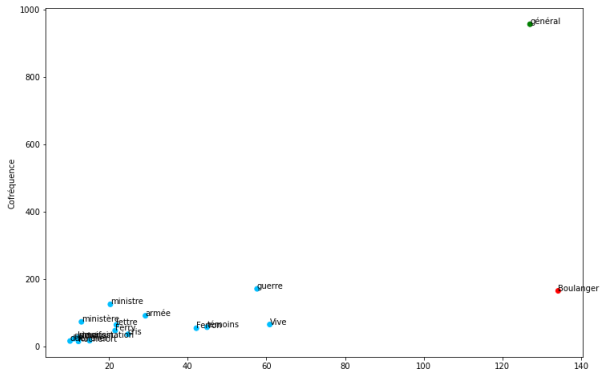


FIGURE 3.14 – Cooccurrents de « Boulanger »

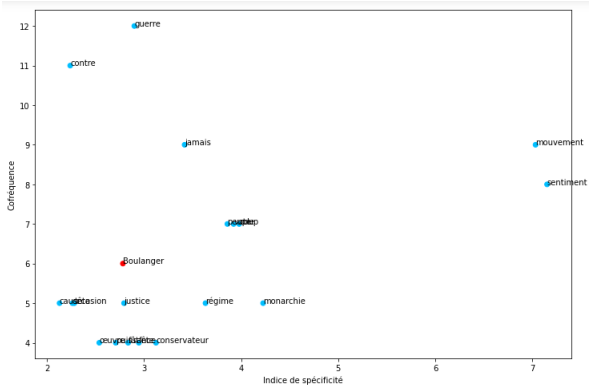


FIGURE 3.15 – Cooccurrents de « populaire »

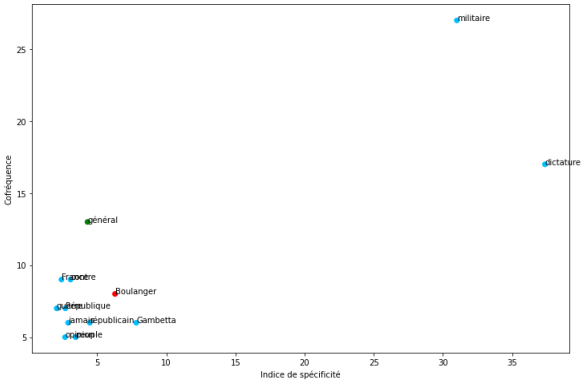


FIGURE 3.16 – Cooccurrents de « dictat. »

On peut constater que, comme pour « populaire », le *token* « Boulanger »obtient un score assez faible, que ce soit en terme de cofréquence ou d'indice de spécificité, avec le *token* « coup Etat », de même que pour « dictat. ». Des résultats assez surprenants (pour le *token* « revanche », non représenté ici, « Boulanger »n'est même pas représenté car ayant un indice inférieur à 2). C'est finalement le graphique des cooccurrents de « Boulanger »qui est le plus significatif, notamment avec le « Vive »qui obtient un indice de spécificité (60.86) assez élevé par rapport à sa cofréquence (65). On peut ainsi estimer que cela montre dans une certaine mesure la popularité du général dont le nom est souvent scandé lors de manifestation (ou lors de la fête nationale), comme le montre également le graphique pour le *token* « Vive »(voir le notebook).

3.5.2 *Word2vec* : recherches de similarités

Word2vec nous permet, par la vectorisation des mots de réaliser, entres autres, des recherches de similarités. Après avoir entraîné un modèle sur notre corpus transformé

en liste de chaînes de caractères et en ayant fait varier la taille du contexte, nous avons ainsi pu dégager quelques résultats. Le modèle se base sur la la similarité cosinus pour mesurer la similarité entre deux vecteurs. Mathématiquement, elle mesure le cosinus de l'angle entre deux vecteurs projetés dans un espace multidimensionnel. La similarité cosinus capte l'angle des vecteurs mots et non leur magnitude. Dans le cadre de la similarité en cosinus, l'absence de similarité (0) est exprimée sous la forme d'un angle de 90 degrés, tandis que la similarité totale de 1 correspond à un angle de 0 degré. En faisant varier la taille du contexte, on peut ainsi s'orienter vers une recherche de similarités syntaxiques (plus elle est petite), ou au contraire une recherche de similitudes thématiques (plus elle est élevée). Nos représentations concernent uniquement des tailles de contexte de 5 et 25, mais nous avons également entraîné un modèle de taille 2 et 50, pour les associations de mots ['Boulanger', 'monarchistes'] et ['bonapartiste','Boulanger'], afin de voir si d'éventuels liens se forment avec la Droite de laquelle il souhaite se rapprocher (discrètement) et se libérer de la tutelle encombrante des radicaux et de la posture de républicain avancé [11].

```
Entrée [42]: model_lematin.wv.most_similar(['bonapartiste','Boulanger'])

Out[42]: [('boulangistes', 0.648832380771637),
          ('modérantisme', 0.6432985067367554),
          ('général', 0.6418592929840088),
          ('Lareinty', 0.6394252777099609),
          ('Ducrot', 0.6379572153091431),
          ('genéral', 0.6297518610954285),
          ('échauffé', 0.6277552247047424),
          ('Appert', 0.6205140352249146),
          ('Lherminier', 0.6190225481987),
          ('Campenon', 0.6165099143981934)]

Entrée [38]: model_lematin.wv.most_similar(['Boulanger', 'monarchistes'])

Out[38]: [('Républiqne', 0.6402148008346558),
          ('opportunistes', 0.6388193964958191),
          ('modérantisme', 0.6346015930175781),
          ('boulangistes', 0.6325921416282654),
          ('radicalisme', 0.63001948595047),
          ('inconséquence', 0.6261537671089172),
          ('intransigeance', 0.6215416789054871),
          ('départager', 0.6110206246376038),
          ('rééditer', 0.6102887392044067),
          ('déchaînées', 0.6063173413276672)]
```

FIGURE 3.17 – Similarités syntaxiques (taille de contexte = 5

```
Entrée [44]: model_lematin_2.wv.most_similar(['bonapartiste','Boulanger'])

Out[44]: [('Lareinty', 0.5612871050834656),
          ('général', 0.5557384490966797),
          ('Schmitz', 0.5548707246780396),
          ('Lalou', 0.541592001914978),
          ('providentiel', 0.534907341003418),
          ('boulangistes', 0.5302249193191528),
          ('boulangiste', 0.521689772605896),
          ('genéral', 0.5053277611732483),
          ('césarisme', 0.5045151710510254),
          ('LoireInférieure', 0.5037518739700317)]

Entrée [46]: model_lematin_2.wv.most_similar(['Boulanger', 'monarchistes'])

Out[46]: [('boulangistes', 0.6313128471374512),
          ('Droite', 0.5813302993774414),
          ('républicains', 0.5753127932548523),
          ('modérantisme', 0.5630757212638855),
          ('radicaux', 0.5609707236289978),
          ('providentiel', 0.5603969097137451),
          ('ExtrêmeGauche', 0.5529002547264099),
          ('SaintArnaud', 0.5474733114242554),
          ('impénitents', 0.5472890138626099),
          ('intransigeance', 0.5353686809539795)]
```

FIGURE 3.18 – Similarités thématiques (taille de contexte = 25)

On peut observer qu'en termes de similarité syntaxique, l'association de « bonapartiste »et « Boulanger »est plus associé aux partisans de ce dernier, ce qui est étonnant au vu de sa fréquence dans le corpus (24 apparitions). Cela peut sans doute signifier que, si le terme le plus similaire ne compte que 24 occurrences, les deux mots ne doivent pas souvent être associés ensemble. C'est d'ailleurs encore moins le cas dans l'approche thématique, qui met davantage en avant Lareinty, contre qui le général s'est battu en duel l'année précédente. A noter tout de même la similarité avec « providentiel »qui, sans être flagrante,

11. B. Joly, *Aux Origines du populisme. Histoire du boulangisme (1886-1890)*..., p. 191.

semble montrer d'une certaine manière la crainte que suscite le général et sa popularité grandissante. L'association « monarchistes »et « Boulanger »donne également des résultats dont les premiers termes ne sont pas extrêmement proche d'elle, que ce soit dans l'approche syntaxique ou thématique. Dans la première, c'est davantage le lexique du républicanisme (République, opportunistes, modérantisme) qui est mis en avant et semble davantage dessiner une opposition, qui ne serait pas étonnante étant donné le froid qu'a jeté l'affaire du duc d'Aumale (tombé sous le coup de la loi du 22 juin 1886 interdisant aux héritiers de la branche royale directe de rester sur le territoire français et de rester dans l'armée, décision que dû appliquer le général).

Conclusion

Il est entendu que nous n'avons fait ici qu'effleurer l'éventail des possibilités qu'offre chacune des approches et des outils utilisés. Malgré tout, bien que l'on ne puisse pas non plus en inférer des conclusions trop hâtives, nous avons pu constater que nous pouvons en dire beaucoup. Notre exploration de *TXM* ou de *spaCy* ne nous a pas permis de trancher définitivement sur le meilleur outil à utiliser à l'avenir, surtout pour traiter un corpus qui va considérablement s'étoffer. Nous pensons, comme nous l'avons déjà dit, que le TAL et la textométrie sont deux approches complémentaires qui offrent des possibilités différentes : retour au texte, partitionnement du texte ou constitution de sous-corpus, accès au lexique plus facile pour *TXM* (sans même évoquer l'analyse factorielle des correspondances ou encore la classification ascendante hiérarchique), tandis que le TAL permet plus de possibilités de manipulation du texte grâce à ses nombreuses librairies *Python*, richement documentées, et de réaliser des calculs (vectorisation des mots et projections avec *word2vec*), de développer des outils basé sur l'apprentissage profond, ou, plus humblement à notre échelle, un dictionnaire de post-correction d'OCR. Nous aurions également aimé utiliser d'autres outils comme Top2vec [12] afin de réaliser du topic modeling, ou encore réaliser une analyse de sentiments (mais il nous semble qu'un texte corrigé sera bien plus intéressant pour mener à bien cette opération).

Nous nous sommes concentrés ici sur un corpus relativement restreint (du moins par rapport à notre ambition de départ, qui est celle du traitement de la presse par le boulangisme) afin de mettre en place une méthodologie adaptée et prendre le temps de découvrir les possibilités qui s'offrent à nous. Cela ne constitue qu'une étape vers un élargissement certain, mais dont la direction reste encore à déterminer : nous hésitons entre notre voie initiale, à savoir le traitement (mais aussi la construction) du boulangisme par la presse (plutôt quotidienne et à fort tirage comme *Le Parisien*, *Le Petit Journal* et *Le Journal*, en évitant la presse clairement orientée d'un bord comme de l'autre (même si l'on a vu avec le directeur du *Matin* montre bien qu'une prétendue neutralité n'est pas gage d'impartialité ou de probité), en élargissant la période étudiée (1886-1890). Mais nous sommes également tenté par une étude plus orientée autour du journal du *Matin*, de sa création en 1884 au début du XXe siècle, période marquée par le tournant de 1897 [13] avec le rachat du journal par Philippe Bunau-Varilla qui le transforme en grand journal populaire, afin d'étudier dans le vocabulaire utilisé, les rubriques proposées, les sujets abordés, le ton employé, ce qui peu expliquer la différence entre *Le Matin ante* et *post* 1897, bien qu'il faille bien faire attention à ne jamais oublier que la seule étude du contenu textuel du journal ne peut suffire et qu'il faut impérativement tenir compte de l'intrication du monde de la (grande) presse avec les puissances politiques et économiques.

12. https://github.com/ddangelov/Top2Vec/tree/master/top2vec
13. https://www.cairn.info/revue-le-mouvement-social-2010-3-page-91.htm

Annexes

Template LaTeX disponible ici : https://github.com/ChartesHN/.

L'ensemble des scripts commentés ainsi que les différents corpus utilisés sont disponibles sur mon github à l'adresse suivante : https://github.com/AdrienEscoda/M1HN

Bibliographie

Awani (Mohammad) et Bassil (Youssef), « OCR post-processing error correction algorithm using google's online spelling suggestion », *Journal of Emerging Trends in Computing and Information Sciences*, vol. 3–1 (2012), URL : http://www.cisjournal.org/journalofcomputing/archive/vol3no1/vol3no1_7.pdf.

Chapuzet (Jean-Charles), *Le Général Boulanger et le boulangisme. Des passions politiques à l'oubli (1886-2005)*, Paris, IEP, 2005, 3 vol.

Chauduri (Arindam), Mandaviya (Krupa), Badelia (Pratixa) et Ghosh (Soumya K.), *Optical Character Recognition Systems for Different Languages with Soft Computing*, Cham, 2017.

Dansette (Adrien), *Le Boulangisme*, Paris, 1938, p. 412.

Garrigues (Jean), « Boulanger ou la fabrique de l'homme providentiel », *Parlement[s], Revue d'histoire politique*–13 (2010/1), URL : https://www.cairn.info/revue-parlements1-2010-1-page-8.html.

— « Le Général Boulanger et le fantasme du coup d'État », *Parlement[s], Revue d'histoire politique*–12 (2009/2), URL : https://www.cairn.info/revue-parlements1-2009-2-page-43.html.

Irvine (William D.), *The Boulanger affair reconsidered : Royalism, Boulangism and the Origins of the Radical Right in France*, New-York, 1989, p. 240.

Joly (Bertrand), *Aux Origines du populisme. Histoire du boulangisme (1886-1890)*, Paris, 2022.

Kalifa (Dominique) et Thérenty (Marie-Ève), « La civilisation du journal. Histoire culturelle et littéraire de la presse française au XIXe siècle », dans dir. Dominique Kalifa, *et al.*, 2011, chap. Ordonner l'information.

Kalifa (Dominique), Thérenty (Marie-Ève) et Vaillant (Alain), « La civilisation du journal. Histoire culturelle et littéraire de la presse française au XIXe siècle », dans dir. Dominique Kalifa, *et al.*, 2011, chap. Le quotidien.

Lebart (Ludovic), Pincemin (Bénédicte) et Poudat (Céline), *Analyse des données textuelles*, Québec, 2019.

McNamara (Daniel), Carolyn (Strange), Wodak (Josh) et Wood (Ian D.), « Mining for the Meanings of a Murder : The Impact of OCR Quality on the Use of Digiti-

zed Historical Newspapers. » *Digital Humanities Quarterly*, vol. 8–1 (2014), URL : https://www.dlib.org/dlib/march09/holley/03holley.html.

NEUDECKER (Clemens), WILMS (Lotte), HACKL (Günter) et BOLLI (Siobhan), *D2.3 Upgraded content*, rapp. tech., Europeana Newspapers Project, 2014, URL : http://www.europeana-newspapers.eu/wp-content/uploads/2015/05/D2.3_Upgraded_content_1.0.pdf.

PINSOLLE (Dominique), *Le Matin (1884-1944). Une presse d'argent et de chantage*, Rennes, 2012.

— « *Le Matin*, les affaires et la politique, 1884-1897 », *Le Mouvement social*–232 (2010/3), URL : https://www-cairn-info/revue-le-mouvement-social-2010-3-page-91.html.

PROVIDENCE (Mathieu), *La contribution paradoxale du boulangisme à l'édification de la démocratie parlementaire*, Paris, Paris X, 2007, 2 vol.

RÉMOND (René), *La Droite en France*, Paris, 1954, p. 544.

ROSE (Holley), « How good can it get ? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs », *D-Lib Magazine*, vol. 15–3/4 (2009), URL : https://www.dlib.org/dlib/march09/holley/03holley.html.

STERNHELL (Zeev), *La Droite révolutionnaire (1885-1914). Les origines françaises du fascisme*, Paris, 1978, p. 441.

The "State of the Art" : A Comparative Analysis of Newspaper Digitization to Date, rapp. tech., ICON, 2015, URL : https://www.crl.edu/sites/default/files/d6/attachments/events/ICON_Report-State_of_Digitization_final.pdf.

WINOCK (Michel), *La Fièvre hexagonale. Les grandes crises politiques (1871-1968)*, Paris, 1986.

— « Populismes français », *Parlement[s], Revue d'histoire politique*–13 (2010/1), URL : https://www-cairn-info/le-xxe-siecle-ideologique-et-politique--9782262042370-page-209.html.

Table des figures

1.1 Les étapes de la reconnaissance optique de caractères Source : CHAUDHURI (Arindam), MANDAVIYA (Krupa), BADELIA (Pratixa), GHOSH (Soumya K.), *Optical Character Recognition Systems for Different Languages with Soft Computing*, Cham, Springer International Publishing, 2017, p. 16. 11

2.1 La première page du quotidien du 13 juillet 1887 19
2.2 La deuxième page du quotidien du 13 juillet 1887 19
2.3 La troisième page du quotidien du 13 juillet 1887 20
2.4 La quatrième et dernière page du quotidien du 13 juillet 1887 20
2.5 Schéma de la méthodologie adoptée afin de traiter notre corpus 22
2.6 Capture d'écran du DataFrame du calcul du CER de nos extraits 25

3.1 Concordancier du corpus pré-traité autour du terme « Boulanger » 32
3.2 Concordancier du corpus brut autour du terme « Boulanger » 32
3.3 Nombre de mots par fichier. Graphique exporté de TXM 33
3.4 Nuage de lemmes réalisé avec *Python* après extraction des données de *TXM* 34
3.5 Fréquence des entités nommées de personnes de 200 occurrences ou plus . . 35
3.6 Nombre d'occurrences du *token* « Boulanger »par quotidien 36
3.7 Nombre de quotidiens contenant n occurrences du *token* « Boulanger » . . 38
3.8 Nombre de quotidiens contenant n occurrences du *token* « Bismarck » . . . 38
3.9 Nombre de quotidiens contenant n occurrences du *token* « Clémenceau » . 39
3.10 Nombre de quotidiens contenant n occurrences du *token* « Goblet » 39
3.11 Nombre de quotidiens contenant n occurrences du *token* « Grévy » 39
3.12 Bigrammes les plus fréquents dans le corpus 41
3.13 Cooccurrents de « coup Etat » 42
3.14 Cooccurrents de « Boulanger » 42
3.15 Cooccurrents de « populaire » 42
3.16 Cooccurrents de « dictat. » 42
3.17 Similarités syntaxiques (taille de contexte = 5 43
3.18 Similarités thématiques (taille de contexte = 25) 43

Table des matières

Résumé **ii**

Remerciements **iii**

Introduction **3**

1 Chaîne de traitement historique - OCR **9**

1.1 Survol historique et descriptif - OCR 9

1.1.1 Brève histoire de l'OCR 9

1.1.2 Définition, fonctionnement et limites de l'OCR 11

1.2 Etat de la numérisation et de l'OCR de la presse ancienne 13

1.2.1 Panorama des différents projets liés à la numérisation et l'océrisation de la presse 13

1.2.2 Facteurs affectant l'OCR de la presse 14

1.3 La nécessité de l'étape du post-traitement 15

2 Le corpus - texte océrisé **18**

2.1 Présentation et récupération du corpus 18

2.1.1 Le corpus : les quotidiens du *Matin* de l'année 1887 18

2.1.2 Présentation du matériau original 18

2.1.3 Disponibilité et méthodes de récupération du corpus 21

2.2 Méthodologie 22

2.2.1 Présentation générale des différentes étapes 22

2.2.2 Pourquoi réocériser avec Tesseract ? 24

2.2.3 Mise en place et comparaison avec la vérité de terrain 26

3 Présentation du jeu de données **27**

3.1 Deux méthodes, plusieurs outils 27

3.1.1 Textométrie et TAL : différences et convergences 27

3.1.2 Les outils 29

3.1.3 *TXM* 29

3.2 Description générale du jeu de données 31
3.2.1 Test sur *TXM* avec le corpus brut et le corpus nettoyé 31
3.2.2 Données générales 33
3.3 Boulanger : une présence importante mais contrastée 34
3.3.1 Le général sous le feu des projecteurs 34
3.3.2 Des apparitions disparates 35
3.4 Une dynamique populiste difficile à mesurer 40
3.5 *NLTK* : recherche des collocations les plus fréquentes 40
3.5.1 Les cooccurrences avec *TXM* 41
3.5.2 *Word2vec* : recherches de similarités 42

Conclusion **47**

Table des figures **53**

Table des matières **54**