Analyse multi-échelle de n-grammes sur 200 années d'archives de presse

THÈSE N^{O} 8180 (2017)

PRÉSENTÉE LE 5 DÉCEMBRE 2017
AU COLLÈGE DES HUMANITÉS
LABORATOIRE D'HUMANITÉS DIGITALES
PROGRAMME DOCTORAL EN MANAGEMENT DE LA TECHNOLOGIE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Vincent Christian BUNTINX

acceptée sur proposition du jury:

Prof. T. A. Weber, président du jury
Prof. F. Kaplan, Dr A. Xanthos, directeurs de thèse
Dr J.-B. Michel, rapporteur
Prof. J. Savoy, rapporteur
Prof. R. West, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE
Suisse
2017

Triste époque que celle où il est plus difficile
de briser un préjugé qu'un atome...
— Albert Einstein

A tous ceux qui m'ont accompagné avec bienveillance et qui m'ont permis d'évoluer.

Remerciements

Un travail de thèse requiert un certain investissement de soi et ce travail n'est finalement jamais totalement le fruit d'une seule et unique personne et ce à plusieurs égards. En premier, les discussions sur le travail en cours avec mes professeurs, directeur et codirecteur de thèse ainsi que mes collègues ont eu bien sûr un impact déterminant. En second mes proches, ma famille et mes amis sont un soutien psychologique important. En dernier, les diverses personnes avec qui j'ai eu des interactions qui m'ont permis de progresser et qui ont nourri mon intellect ainsi que ma personnalité durant mon existence jusqu'à maintenant.

Ainsi, je souhaite vivement remercier les personnes suivantes pour leur soutien, leur bienveillance et leur rôle déterminant sur ce travail de thèse :

- Mon directeur de thèse, **Frédéric Kaplan**, qui m'a guidé avec perspicacité, intelligence et bienveillance tout au long de cette thèse.
- Mon codirecteur de thèse, **Aris Xanthos**, qui m'a fourni de précieux conseils et beaucoup de rigueur avec une égale bienveillance.
- Mon compagnon **Anthony Van Rossem**.
- Mes parents, **Brigitte Decoster** et **Danny Buntinx**.
- Ma soeur, **Virginie Buntinx** et son mari **Virgile Bouillon**.
- Mes grands parents **Irène Devogeleer**†, **Victor Decoster** et **Josée De Breucker**.
- Ma famille proche, **Gerda Decoster**, **René Vanbuyten**†, **Danielle Decoster**†, **Jean-François Lefrancq**, **Antoine Lefrancq** et **Jérôme Lefrancq**.
- Mes amis proches, **Jonathan Page**, **Johann Breitenhuber**, **Sebastien Jordan**, **Marc Wuestenberghs** et en particulier **Patricia Marlet** et **Patrick Staeger**.
- Mes senseis et amis proches, **Hedwig Blancke**, **Jean-Marc Spothelfer** et **Antonio Stifani**.
- Tous mes collègues du DHLAB et particulièrement **Alicia Foucart**.

Ainsi que les nombreuses autres personnes qui m'ont accompagné ou simplement traversé ma vie en y ajoutant du positif.

Lausanne, 27 Octobre 2017 V. B.

Abstract

The recent availability of large corpora of digitized texts over several centuries opens the way to new forms of studies on the evolution of languages. In this thesis, we study a corpus of 4 million press articles covering a period of 200 years. The thesis tries to measure the evolution of written French on this period at the level of words and expressions, but also in a more global way by attempting to define integrated measures of linguistic evolution.

The methodological choice is to introduce a minimum of linguistic hypotheses in this study by developing new measures around the simple notion of n-gram, a sequence of n consecutive words. The thesis explores on this basis the potential of already known concepts as temporal frequency profiles and their diachronic correlations, but also introduces new abstractions such as the notion of resilient linguistic kernel or the decomposition of profiles into solidified expressions according to simple statistical models. Through the use of distributed computational techniques, it develops methods to test the relevance of these concepts on a large amount of textual data and thus allows to propose a virtual observatory of the diachronic evolutions associated with a given corpus.

On this basis, the thesis explores more precisely the multi-scale dimension of linguistic phenomena by considering how standardized measures evolve when applied to increasingly long n-grams. The discrete and continuous scale from the isolated entities ($n = 1$) to the increasingly complex and structured expressions ($1 < n < 10$) offers a transversal axis of study to the classical differentiations that ordinarily structure linguistics : syntax, semantics, pragmatics, and so on. The thesis explores the quantitative and qualitative diversity of phenomena at these different scales of language and develops a novel approach by proposing multi-scale measurements and formalizations, with the aim of characterizing more fundamental structural aspects of the studied phenomena.

Keywords : big data, corpus analysis, frequency profile, linguistic distance, linguistic evolution, n-grams analysis, press corpus, resilient kernel, word resilience.

Résumé

La récente disponibilité de grands corpus de textes numérisés s'étalant sur plusieurs siècles ouvre la voie à de nouvelles formes d'études sur l'évolution des langues. Dans cette thèse, nous étudions un corpus de 4 millions d'articles de presse s'étalant sur une période de 200 ans. La thèse tente de mesurer l'évolution du français écrit sur cette période à la fois au niveau des mots et des expressions, mais aussi de manière plus globale en tentant de définir des mesures intégrées de l'évolution linguistique.

Le choix méthodologique est d'introduire un minimum d'hypothèses linguistiques dans cette étude en développant de nouvelles mesures autour de la notion simple de n-gramme, une séquence de n mots consécutifs. La thèse explore sur cette base, le potentiel de concepts déjà connus comme les profils fréquentiels temporels et leurs corrélations diachroniques, mais introduit également de nouvelles abstractions comme la notion de noyau linguistique résilient ou la décomposition de profils en expressions solidifiées suivant des modèles statistiques simples. Elle développe, grâce à l'utilisation de techniques distribuées de calculs, des méthodes pour tester la pertinence de ces concepts sur une grande quantité de données textuelles et permet ainsi de proposer un véritable observatoire virtuel des évolutions diachroniques associées à un corpus donné.

Sur cette base, la thèse explore plus précisément la dimension multi-échelle des phénomènes linguistiques en considérant la manière dont des mesures standardisées évoluent quand elles sont appliquées à des n-grammes de plus en plus en longs. L'échelle discrète et continue partant des entités isolées ($n = 1$) jusqu'aux expressions de plus en plus complexes et structurées ($1 < n < 10$) offre un axe d'étude transverse aux différenciations classiques qui d'ordinaire structurent la linguistique : la syntaxe, la sémantique, la pragmatique, etc. La thèse explore la diversité quantitative et qualitative des phénomènes à ces différentes échelles de la langue et développe une approche inédite en proposant des mesures et formalisations multi-échelles, ayant pour objectif de caractériser des aspects structurels plus fondamentaux des phénomènes étudiés.

Mots clés : analyse de corpus, analyse des n-grammes, big data, corpus de presse, distance linguistique, évolution linguistique, noyau résilient, profil fréquentiel, résilience de mots.

Table des matières

Table des figures

Liste des tableaux

Introduction Partie I

1 Enjeux scientifiques et culturels

1.1 Comment le big data transforme la science

Les avancées technologiques et scientifiques transforment continuellement notre environnement et l'évolution rapide des technologies de l'information a entrainé des mutations profondes de nos sociétés. Nos capacités de stockage et traitement augmentent à grande vitesse, et de plus en plus de données deviennent disponibles sous forme numérique. Les mégadonnées ou le big data, ensembles de données extrêmement volumineux, font leur apparition et ont contraint les chercheurs et scientifiques à inventer de nouvelles techniques pour les traiter. Aujourd'hui, les perspectives du big data sont prometteuses et beaucoup restent encore à explorer. Cette explosion quantitative de données sous forme numérique a permis de nouvelles façons d'analyser et visualiser le monde réel.

Aujourd'hui, le big data est en passe de métamorphoser le domaine scientifique ainsi que de nombreux secteurs de l'industrie. C'est notamment le cas, pour le secteur de la santé qui concentre de nombreuses bases de données complexes afin de résoudre des problèmes difficiles comme, par exemple, la prévision des risques de cancers ou tumeurs sur la base de milliers de variables différentes collectées sur des sujets suivis durant des dizaines d'années. Les gains potentiels sont nombreux : médecine génomique et personnalisée, gestion des populations, campagnes de dépistage mieux ciblées, etc. Ainsi, le phénomène d'explosion des données dans le domaine de la bioinformatique (séquence génomique, images de structures physiologiques, etc) (Greene *et al.*, 2014) a été suivie de près par celle des humanités digitales, de plus en plus de données anciennes et archives revêtant diverses formes étant numérisées afin de reconstruire le passé de l'humanité (Kaplan et di Lenardo, 2017).

Des compagnies comme Google, Facebook, Twitter, Instagram, Whatsapp ou autres collectent de plus en plus d'informations sur nos interactions avec le web, sur nos déplacements et sur nos agissements en général (même en dehors de la toile). Le monde qui nous entoure est également de plus en plus connecté, générant de nombreuses données numériques supplémentaires. De plus en plus d'acteurs culturels comprennent le rôle décisif de leurs archives et les numérisent afin de les valoriser au travers de nouveaux outils numériques.

Cette avalanche de données et les techniques big data ont donné naissance à un débat au travers duquel sont mises en cause les traditionnelles méthodes scientifiques et particulièrement la nécessité de passer par l'abstraction de la modélisation afin de résoudre les problèmes complexes alors que le big data permet d'outrepasser cette étape.

Dès lors, plusieurs chercheurs se sont interrogés sur la nécessité de passer par cette abstraction approximativement correcte, mais formellement toujours erronée de la réalité. En 2008, un article alimentant ces débats est publié dans Wired (Anderson, 2008), citant des exemples de réussites de Google à traiter divers problèmes difficiles sans le besoin de modélisation, mais avec le soutien de grandes bases de données. L'auteur argumente que, de plus en plus, les modèles théoriques deviennent facultatifs face à la quantité de données que nous pouvons traiter et il note que plus de données ne sont pas seulement "plus", mais que cela peut tout changer dans la façon d'appréhender et résoudre les problèmes. Il cite notament Peter Norvig, le directeur de recherche de Google : " All models are wrong, and increasingly you can succeed without them" (Tous les modèles sont faux, et de plus en plus vous réussirez sans eux), mettant à jour la célèbre phrase du statisticien George Box (Box, 1976) : "All models are wrong, but some are useful" (Tous les modèles sont faux, mais certains sont utiles).

Sans entrer dans un tel débat, il est clair que le big data à l'avantage de permettre la création de méthodes robustes, quasiment indépendantes de la nature des données, réduisant les hypothèses nécessaires en comparaison aux méthodes classiques sous-tendues par des modélisations théoriques plus poussées et plus rigides. Ainsi, ces nouvelles méthodes créées sont génériques, réutilisables sur d'autres bases de données et indépendantes de nombreux paramètres de modélisation.

Dans cette thèse, nous utilisons ce type de méthodologie afin d'étudier des problèmes complexes tout en minimisant les hypothèses de modélisation nécessaires et nous tentons de découvrir jusqu'à quel point ces méthodes peuvent nous aider à comprendre les mécanismes principaux sous-jacents aux problèmes étudiés.

1.2 Les grandes bases de données textuelles vont transformer la linguistique de corpus

La linguistique est le domaine scientifique dont l'objet d'étude est le langage. Parmi les premiers linguistes a avoir contribué à la description formelle et scientifique du langage et des langues, Ferdinand de Saussure développe une approche structuraliste notamment au travers de son cours de linguistique générale (Saussure, 1916). Cet ouvrage, publié après sa mort par certains de ses étudiants qui voulaient lui rendre hommage, est devenu un classique dans ce domaine. Il a contribué à imposer la linguistique structurale au travers d'un profond travail de réflexion et formalisation sur les principes fondamentaux du langage.

Selon la vision structurale, la langue est vue comme la partie sociale du langage. La faculté de communiquer à travers le langage se décompose alors en une partie individuelle, le discours,

et une partie collective et sociale, la langue. La langue est donc l'outil collectif permettant aux individus d'interpréter et de formuler des idées à partir de signes (signes vocaux ou signes écrits). Le langage est défini comme la capacité des êtres humains à communiquer au moyen d'un système de signes et la langue est l'outil social permettant aux individus de communiquer entre eux. Par exemple, si un homme ne maîtrise que la langue anglaise et essaie de communiquer avec un autre qui ne maîtrise que la langue française, il est peu probable qu'ils se comprennent.

Le corpus est un ensemble de textes, documents, images ou vidéos qui constituent un panel de données issues du monde réel. Les corpus permettent, en autres, l'étude scientifique des objets dont ils sont issus comme la langue dans le cas de corpus textuels. Si une opposition a existé entre la linguistique "sans corpus" et celle "avec corpus", il n'est pas un linguiste qui ne manipule pas aujourd'hui un tel objet (Mayaffre, 2005). Un des arguments qui, dans le passé, pouvait conduire un linguiste à écarter l'étude de corpus était son "impureté".

En effet, le corpus est inévitablement "contaminé" par divers bruits vu qu'il provient d'un discours réel qui se manifeste forcément au travers de l'histoire, de la psychologie et de la société. Les études de corpus sont donc parfois placées sous le terme de la psycho-linguistique ou le socio-linguistique. Toutefois, l'utilisation du langage au jour le jour produit typiquement des données textuelles écrites et l'analyse de telles données ne peut que s'avérer utile à la compréhension de la langue. Divers types de corpus existent et il est important de tenir compte de leurs spécificités afin de ne pas surgénéraliser les résultats de l'analyse.

Cependant, l'un des problèmes courants de l'analyse de corpus est que celui-ci ne constitue que rarement un échantillon représentatif (Biber, 1993) de la langue. Bien sur, même avec peu de représentativité, cela ne signifie pas qu'aucune information linguistique ne peut être extraite de celui-ci. D'un autre coté, l'abondance actuelle de données numériques et l'accès facilité à de grandes collections de textes numériques a permis le développement du domaine de la linguistique de corpus. Cette linguistique appliquée se base sur des données réelles et a donc l'avantage de l'empirisme. L'objectif est alors de détecter les patterns et régularités afin d'extraire des connaissances linguistiques d'un corpus. Il est donc possible dans certains cas d'émettre des hypothèses sur la langue afin de donner de nouveaux points de vue en se basant sur ces ensembles de textes du monde réel, étudiant la signification dans le discours et les interactions sociales produites à l'aide de la langue, mais en considérant aussi divers bruits dus à la spécificité de ces corpus.

Toutefois, comme pour de nombreuses disciplines, l'avènement du big data est en train de changer profondément le domaine de la linguistique de corpus. En effet, si les corpus constituent des échantillons souvent non représentatifs de la langue, ceux-ci deviennent de plus en plus grands et peuvent prétendre à une représentativité augmentée dans ce contexte d'explosion de données. Il est donc important de continuer à développer des méthodes à grande échelle afin de traiter indépendamment divers corpus larges tout en minimisant les modélisations théoriques permettant d'extraire des informations utiles sur la langue.

Ainsi, ces méthodes permettent une automatisation sans précédent et peuvent aider les linguistes à extraire l'information pertinente de ces corpus afin de valider des hypothèses sur la langue. Dans cette thèse, nous allons développer des méthodes avec pour objectif : la robustesse, l'indépendance vis-à-vis du type de données traitées et l'extraction du contenu linguistique de grands corpus textuels en ciblant spécifiquement la langue plutôt que le bruit inhérent du corpus.

1.3 De nouvelles méthodes pour l'étude de l'évolution des langues

La linguistique diachronique est un domaine de la linguistique dont l'objet d'étude est l'évolution des langues. Celle-ci évoluent continuellement (McMahon, 1994; Labov, 1994). Certains mots sont abandonnés, de nouveaux mots apparaissent en identifiant de nouvelles significations, concepts et objets, remplaçant parfois d'autres mots (Steels et Kaplan, 1998). La rencontre de deux populations linguistiques peut entraîner un mélange de différentes caractéristiques des deux langues (Arends *et al.*, 1994). Diverses controverses et consensus existent (Christiansen et Kirby, 2003) sur l'évolution du langage. Parmi ces consensus, figure l'image émergente et complexe de l'évolution du langage à travers les interactions de trois systèmes adaptatifs différents : l'apprentissage individuel, la transmission culturelle et l'évolution biologique. L'évolution du langage est ainsi définie par l'évolution même de ces trois systèmes qui interagissent continuellement, le langage lui-même étant le résultat issu de ces interactions. L'une des controverses est la nature du rôle de l'évolution biologique par rapport à celle de l'évolution culturelle. Ces questions conduisent inévitablement à la question complexe de l'origine du langage où le consensus est encore plus rare.

L'étude du langage a suscité un certain intérêt dans le domaine de l'informatique ces dernières années, ces chercheurs ont testé des théories sur l'origine et l'émergence de la langue (Ke et Holland, 2006; Bartlett et Kazakov, 2005), l'évolution culturelle de la langue (Kirby, 2001; Wang *et al.*, 2004; Steels, 2011) ou son évolution dans un contexte de deux différentes espèces interagissant (Kosmidis *et al.*, 2005) principalement au travers de modèles basés sur des agents multiples et des simulations Monte Carlo. Ces simulations ont l'avantage d'être des expériences reproductibles, cependant elle manquent généralement de réalisme dans l'état initial de la simulation. L'utilisation du terme "évolution" fait également débat posant la question "Les langues évoluent-elles ou changent-elles simplement ?". Des arguments comparant l'évolution de la langue (évolution non biologique) et l'évolution biologique sont articulés (Steels, 2016) en faveur de l'utilisation du terme "évolution" car le phénomène d'évolution de la langue possède des caractéristiques similaires à celles d'une évolution biologique classique.

L'évolution de la langue peut être mesurée en étudiant les corpus diachroniques. Au cours des dernières années, de plus en plus de données textuelles et de grands corpus offrent l'occasion d'analyser l'évolution linguistique aussi à travers le prisme d'un grand corpus de données textuelles (Dipper, 2008; Bender et Good, 2010; Weikum *et al.*, 2012). Des chercheurs ont constitué un immense corpus formé des livres numérisés par Google, les Google Books, repré-

sentant environ 4% de tous les livres jamais publiés afin d'en étudier les grandes tendances culturelles et linguistiques, faisant émerger le domaine nommé Culturomics (Michel *et al.*, 2011). L'étude de ces corpus a permis d'inférer des hypothèses à propos de l'évolution de la culture, de l'histoire ou de la langue au travers du passé.

Cette nouvelle discipline a montré un grand potentiel pour générer ce type d'hypothèses sur la base de données textuelles réelles. En effet, le corpus de Google Books est par nature bien plus représentatif de la langue que la plupart des autres corpus de taille réduite en terme de quantité de données. Cependant, il a le désavantage d'être imparfait de par l'évolution de sa composition, son degré d'hétérogénéité évoluant avec le temps (Pechenick *et al.*, 2015a). Un exemple typique étant l'introduction de textes purement scientifiques dès 1900 pouvant partiellement fausser l'analyse diachronique. Ainsi, il est important que le corpus puisse être gardé sous contrôle tout en possédant un contenu bien défini.

Dans cette thèse, nous utilisons principalement un corpus de presse de 4 millions d'articles, bien défini, mais imparfait à plusieurs niveaux. Nous tentons donc de développer des méthodes qui soient robustes vis-à-vis de ces diverses évolutions (comme celle de l'hétérogénéité de la composition du corpus) afin de ne pas confondre l'évolution de la langue avec l'évolution de phénomènes non linguistiques (comme, par exemple, l'évolution de la diversité des sujets).

1.4 Les raisons pour étudier l'évolution de la langue aujourd'hui

Au cours de la dernière décennie, de plus en plus de sources a priori inattendues affectent notre expression linguistique. Aujourd'hui, nous nous exprimons continuellement au travers de nouvelles formes numériques d'expressions textuelles comme les mails, SMS, logiciels de traitement de texte, moteurs de recherche, traducteurs automatiques, articles de blogs ou wikipédia. Cela affecte non seulement notre syntaxe et lexique, mais aussi nos habitudes cognitives (Carr, 2011). Une grande variété de processus algorithmiques fonctionnent comme des intermédiaires dans les chaînes textuelles exprimées, en transformant les textes initiaux en d'autres textes. Les algorithmes interviennent également sur notre expression textuelle grâce à des services de suggestions ou de correction orthographique. D'autres algorithmes, comme les bots éditeurs de texte vont même jusqu'à produire du texte directement par eux-mêmes.

Deux hypothèses concernant l'effet de la médiation algorithmique ont été formulées (Kaplan, 2014). La première est que le changement linguistique global de la langue est dans une phase d'accélération anormale par rapport à son évolution naturelle en raison de l'intervention systématique des algorithmes sur notre expression linguistique. La seconde est que cette transformation linguistique se caractérise par un processus de réduction de la diversité, en raison des algorithmes forçant les expressions irrégulières à devenir statistiquement plus régulières pour des motifs économiques (par exemple la publicité sur le moteur de recherche Google). Ces hypothèses n'ont pas été testées sur des données linguistiques réelles et l'article conclut que de nouveaux outils doivent être construits afin de comprendre et mesurer cette évolution linguistique globale.

Nous considérons donc les éléments suivants :
– Le besoin crucial de nouvelles méthodes de mesure de l'évolution de la langue.
– L'avènement du big data transformant de nombreux domaines scientifiques.
– Les développements récents du domaine de la linguistique de corpus.
– L'émergence des études diachroniques sur de plus en plus grands corpus textuels.

A travers ces éléments, se précise une question fondamentale : "Que pouvons nous dire sur l'évolution de la langue via des méthodes de mesure de type big data, automatiques et minimisant les hypothèses de modélisation, appliquées sur de grands corpus textuels diachroniques ?".

C'est autour de cette question générale que s'articule la plupart des analyses effectuées dans ce travail de thèse. Pour cela, nous avons besoin de développer des nouvelles méthodes et les tester sur un corpus bien délimité. Dans cette thèse, nous utilisons un grand corpus de presse représentant 4 millions d'articles couvrant 200 années d'archives. Il est subdivisé en deux journaux distincts, la Gazette de Lausanne (GDL) et le Journal de Genève (JDG).

Sur ce corpus, nous allons effectuer une analyse linguistique diachronique non pas au sens de la linguistique classique, mais plutôt en se plaçant dans le cadre particulier d'études big data. Ainsi, chaque fois que nous utiliserons le terme "analyse linguistique diachronique" dans ce travail de thèse, cela signifiera simplement que nous analysons un corpus diachronique avec pour objectif d'en extraire des informations sur l'évolution linguistique de celui-ci. Ces informations seront éventuellement généralisables à la langue elle-même dans certains cas particuliers, conditionnellement au fait de reproduire ces même résultats sur d'autres corpus écrits dans une même langue.

2 Enjeux méthodologiques

2.1 Représentativité d'un corpus pour étudier la langue

L'objectif est d'extraire de ce corpus de presse des informations sur la langue française. Il est clair qu'à plusieurs égards, ce corpus ne représente pas l'ensemble de la langue. En effet, ce corpus peut être vu comme un échantillon dont la représentativité est à discuter. En outre, sa représentativité peut dépendre du type de méthodes que nous utilisons dans ce travail.

D'un autre coté, même l'ensemble de toute la presse ne représente pas la langue dans sa globalité puisque qu'il s'agit d'un cas spécifique de l'utilisation de la langue. Elle aussi, peut être vue comme un échantillon de la langue, mais pas comme un échantillon sans biais. Il est donc nécessaire d'adopter une vision méthodologique permettant de les prendre en compte.

Nous utilisons un raisonnement ensembliste simple pour représenter ce corpus de presse et son lien avec la langue. Soit l'ensemble O de tous les mots uniques prononcés en français dans le monde durant une période donnée (par exemple une année). Soit l'ensemble E de tous les mots uniques écrits en français dans le monde durant cette même période donnée. D'un point de vue intuitif, l'ensemble E est contenu entièrement dans l'ensemble O.

Cependant d'un point de vue mathématique, rien n'empêche que ces deux ensembles aient une partie disjointe, et ce particulièrement si la période considérée est courte. Toutefois, nous faisons l'hypothèse que l'ensemble E est totalement contenu dans O sans aucune partie disjointe. L'ensemble E contient tous les mots uniques provenant de tous les corpus de textes écrits quels qu'ils soient. Nous définissons l'ensemble P comme étant l'ensemble de tous les mots uniques qui ont été écrits dans la presse en français durant la période fixée précédemment. Cette fois, l'ensemble P est mathématiquement contenu dans l'ensemble E.

Cet ensemble constitue un échantillon de la langue, mais pas sans biais, car il est orienté par le contexte particulier de l'utilisation de la langue française pour communiquer des nouvelles via des articles de presse. Nous pouvons subdiviser l'ensemble P en représentant les mots uniques écrits en français dans des corpus de presse particuliers comme ceux de GDL et JDG.

Dans une période donnée, les différents corpus de presse contiennent des mots similaires ainsi que des mots qui ne se retrouvent pas dans l'autre corpus. Ils possèdent donc chacun une partie disjointe les uns des autres. Les ensemble GDL et JDG représentent chacun l'ensemble de tous les mots uniques qui ont été écrits dans le journal de GDL, respectivement le journal de JDG, en français durant la période fixée précédemment.

La représentation ensembliste simple constituée des ensembles O, E, P, GDL et JDG précédemment définis est présentée dans la Figure 2.1.

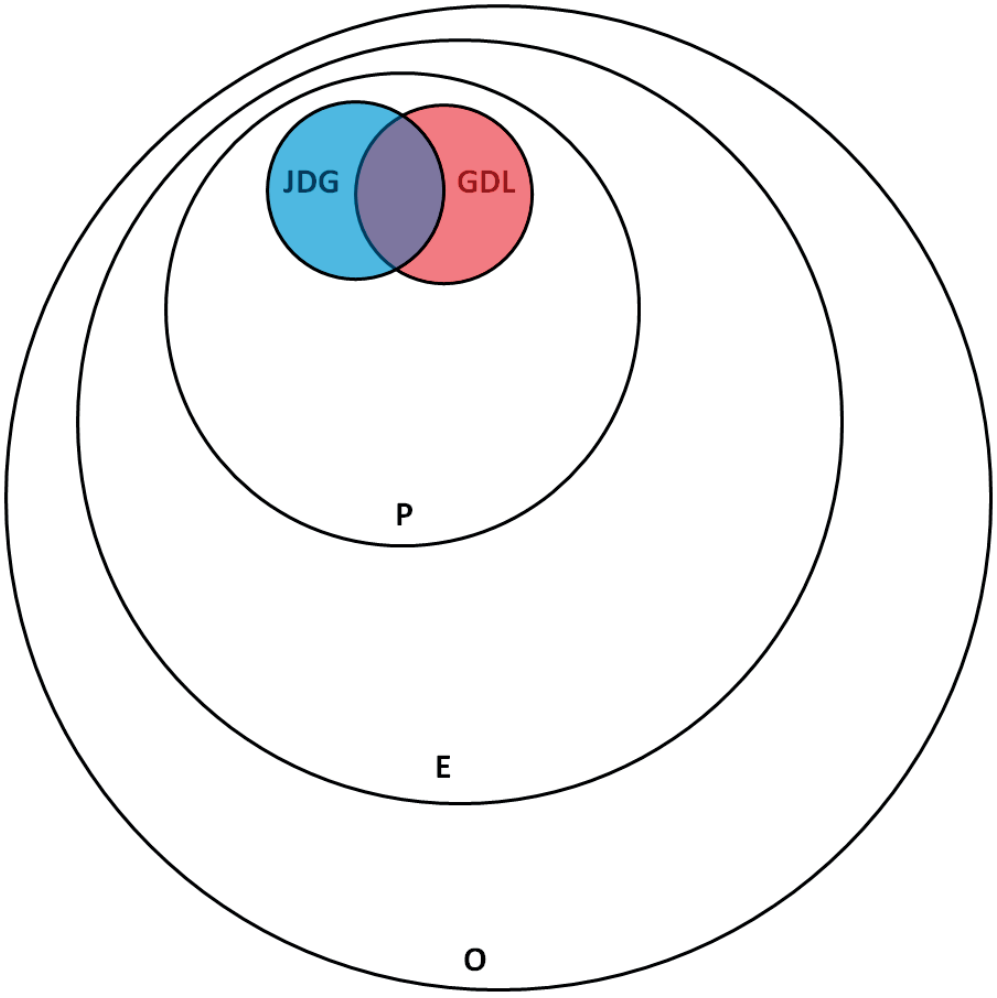


FIGURE 2.1 – Représentation des ensembles O (ensemble de tous les mots uniques prononcés en français dans le monde durant une période T), E (ensemble de tous les mots uniques écrits en français dans le monde durant une période T), P (ensemble de tous les mots uniques écrits en français dans la presse durant une période T), GDL (ensemble de tous les mots uniques écrits en français dans le journal GDL durant une période T) et JDG (ensemble de tous les mots uniques écrits en français dans le journal JDG durant une période T)

Un raisonnement simple nous permet de penser que certains mots sont plus fondamentaux dans le fonctionnement de la langue que d'autres. Saussure faisait notamment référence aux éléments linguistiques internes et externes (Saussure, 1916). Les éléments externes sont généralement issus des relations entre la langue et l'histoire, la politique, les institutions, la géographie, etc. Les éléments internes font partie d'un système propre à la langue et sont considérés stables. Cette vision duale est centrale dans ce travail de thèse, car nous étudions la langue au travers d'un grand corpus de presse et que celui-ci introduit un grand nombre d'éléments linguistiques externes. Ces éléments sont intéressants à étudier, car ils retracent l'histoire, la culture, la politique et toutes les informations généralement couvertes par la presse, mais nous discuterons aussi de la possibilité d'étudier les éléments internes de la langue qui se rapportent plus à son fonctionnement en tant que système.

Nous remarquons intuitivement que les mots uniques écrits dans JDG, mais pas dans GDL ainsi que les mots uniques écrits dans GDL, mais pas dans JDG ont une probabilité élevée de faire partie des éléments externes (dépendant de la longueur de la période T considérée). Toutefois, les éléments écrits dans les deux journaux au cours de cette même période T ne sont pas pour autant des éléments internes, car ceux-ci peuvent faire référence ensemble à des éléments externes à la langue et notamment au travers de l'histoire. Bien que cette représentation implique la langue française, elle est indépendante de la langue de référence et peut être appliquée quelle que soit la langue. Tous ces mots uniques ne sont pas utilisés aussi fréquemment et il est intéressant d'étudier la fréquence d'utilisation de ces mots, en particulier dans un contexte diachronique.

Les mots sont des signes au sens de la linguistique structurale, mais un signe n'est pas une étiquette arbitraire qui dénote un concept ou une représentation mentale comme on serait tenté de se la représenter, par exemple via un dictionnaire, car elle serait réduite à une simple nomenclature. Le signe est une entité double qui établit le lien entre un concept et une image acoustique. C'est une relation entre la pensée et le son. Ce lien est arbitraire, car il repose sur des conventions. Le concept s'appelle le signifié, l'image acoustique s'appelle le signifiant.

Il est intéressant de noter que le sens d'un signe dépend de la relation entre tous les signes du système et non d'un contenu intrinsèque du signe. En effet, dans le cas du signifiant, ce n'est pas le son d'un mot qui construit l'image acoustique, mais les différences phoniques qui le distinguent des autres. Dans le cas de signifié, nous déterminons le sens en fonction de la différence avec les concepts existant dans la langue. A titre d'exemple, si nous divisons le spectre lumineux selon les couleurs du spectre continu visible, un nombre discret de mots déterminera le concept de couleurs. Mais en comparant différentes langues, les mots peuvent se référer à différentes nuances de couleurs et ne représentent donc pas exactement le même concept. En outre, le langage peut adopter un nombre de mots différents pour la représentation de la couleur visible de sorte que la précision n'est pas la même.

Le sens est donc donné par la relation entre le signifié et le signifiant et la langue donne un sens aux mots selon la façon dont elle subdivise le réel. Saussure définit d'ailleurs la valeur

linguistique d'un signe comme la relation entre le signe et tous les autres signes partageant des unités de sens communes. La valeur du signe émerge donc par les autres signes dans une relation d'identité et d'opposition. La notion de valeur linguistique est donc déterminée par l'analyse du rapport entre les éléments du système de la langue.

Ainsi la représentation de la langue en terme lexical est intéressante, mais elle n'est en réalité qu'une approximation et ne tient pas compte des innombrables combinaisons possibles de ces mots en vue de créer des phrases syntaxiquement et sémantiquement correctes. En effet, la langue peut être représentée comme l'ensemble de tous les textes dont chacun possède sa propre temporalité et est composé d'une séquence de phrases, elles-mêmes composées de séquences de mots. Ainsi, une modélisation plus complète de la langue consiste à considérer également les combinaisons de mots. Au cours de cette thèse, nous développerons des méthodes basées sur l'étude diachronique de ces combinaisons appelées les n-grammes.

2.2 Etudes à partir des n-grammes

Le suffixe gramme a pour origine grámma en grec et signifie un signe ou un écrit. En linguistique, le terme "n-gramme" désigne une séquence de mots dont le nombre total vaut n. Il suffit donc de remplacer *n* par un entier pour préciser la taille de la séquence.

La définition formelle du n-gramme est simple :

Définition 1. *Un n-gramme est une série ordonnée de n mots consécutifs.*

Remarquons que "séquence" est équivalent à "série ordonnée", mais ce dernier à l'avantage de mettre en évidence l'importance de l'ordre des mots. A titre d'exemple, dans la phrase "Je suis heureux", les mots "Je", "suis" et "heureux" sont des 1-grammes, les entités "je suis" et "suis heureux" sont des 2-grammes, tandis que "je suis heureux", la phrase elle-même, est un 3-gramme. Notons que le terme n-gramme peut aussi référer dans la littérature à une combinaison de caractères. Dans ce cas, la chaîne "je suis" a sept 1-gramme ("j", "e", " ", "s", "u", "i" et "s"), six 2-grammes ("je", "e ", " s", "su", "ui" et "is"), cinq 3-grammes ("je ", "e s", " su", "sui" et "uis"), quatre 4-grammes ("je s", "e su", " sui" et "suis"), trois 5-grammes ("je su", "e sui" et " suis"), deux 6-grammes ("je sui" et "e suis") et un 7-gramme ("je suis"). Dans cette thèse le terme "n-gramme" se réfère à un n-gramme de mots, donc une séquence de *n* mots, tandis que le terme "n-gramme de caractères" se réfère à une séquence de *n* caractères.

Les analyses de n-grammes sont étudiés depuis longtemps en linguistique computationnelle et le concept est utilisé dans de nombreux domaines, tels que la traduction statistique et automatique qui exploite des techniques d'apprentissage automatique afin de construire une modélisation de la langue permettant non seulement d'identifier la langue utilisée (Zissman et Singer, 1994; Hakkinen et Tian, 2001; Murthy et Kumar, 2006; Ljubesic *et al.*, 2007; Tomović et Janičić, 2007; Rehurek et Kolkus, 2009; Tromp et Pechenizkiy, 2011; Takçı et Güngör, 2012; Buck *et al.*, 2014), mais aussi d'effectuer des traductions automatiques tout en étudiant le problème

difficile de son évaluation (Niesler et Woodland, 1996; Doddington, 2002; Papineni *et al.*, 2002; Culy et Riehemann, 2003; Zens et Ney, 2006; Federico et Cettolo, 2007; Pauls et Klein, 2011; Cho *et al.*, 2014; Durrani *et al.*, 2015). L'évaluation de ces modèles s'effectue souvent au travers de la construction et l'amélioration de divers scores (typiquement le score nommé BLEU, BiLingual Evaluation Understudy) sur la base de modèles statistiques existants de la langue.

Les domaines de l'identification et la modélisation de la langue ainsi que la traduction automatique sont aussi liés à la détection des expressions multi-mots. En effet, la plupart des traducteurs automatiques sont généralement moins efficaces concernant les expressions multi-mots et ce domaine se développe indépendamment (Remaki et Meunier, 2000; Biskri *et al.*, 2004; Blei et Lafferty, 2009; Ramisch *et al.*, 2010; Constant *et al.*, 2012; Lyse et Andersen, 2012; Tsvetkov et Wintner, 2012; Emms et Jayapal, 2014; Salehi *et al.*, 2015) tout en permettant d'affiner les modèles de traduction automatique.

En outre, la détection des expressions multi-mots est également liée à l'extraction automatique de mots clés et de résumés de textes (Hardy *et al.*, 2002; Hulth, 2003; Lin et Hovy, 2003; Banko et Vanderwende, 2004; Erkan et Radev, 2004; Lin, 2004; Giannakopoulos *et al.*, 2008; Kim *et al.*, 2010; Yang *et al.*, 2015) ainsi qu'à la détection des événements (Hamid *et al.*, 2006; Yuan *et al.*, 2006; Mcquiggan *et al.*, 2007; Snowsill *et al.*, 2010a; Lampos et Cristianini, 2012; Snowsill *et al.*, 2010b; Bettadapura *et al.*, 2013; Atefeh et Khreich, 2015; Olteanu *et al.*, 2015) dans ces même textes. Le domaine de la tokenization ainsi que les études de correction de l'OCR (Nagao et Mori, 1994; Tong et Evans, 1996; Torres-Carrasquillo *et al.*, 2002; Mcnamee et Mayfield, 2004; McNamee, 2008; Bassil et Alwani, 2012b,a; Evershed et Fitch, 2014; Kumar, 2016) se servent plus souvent de la notion d'n-grammes de caractères et ces études sont particulièrement intéressantes dans le sens qu'elles décrivent typiquement des techniques de prétraitement de données permettant d'améliorer leur qualité et donc les résultats d'autres types d'analyses appliqués à ces données.

Le domaine de l'identification d'auteurs permet de retrouver des régularités statistiques dans l'écriture des textes afin de les classer et identifier son auteur (Stamatatos *et al.*, 2000; Abou-Assaleh *et al.*, 2004; Stamatatos *et al.*, 2006; Jardino, 2006; Barrón-Cedeño et Rosso, 2009; Koppel *et al.*, 2009; Stamatatos, 2009; Sidorov *et al.*, 2013; Peng *et al.*, 2016). Divers autres problèmes sont similaires à celui de l'identification d'auteurs comme l'analyse des sujets et la classification de textes (Cavnar *et al.*, 1994; Damashek, 1995; Gildea et Hofmann, 1999; Jalam et Chauchat, 2002; Greevy et Smeaton, 2004; Wang *et al.*, 2007; Wei *et al.*, 2008; O'Connor *et al.*, 2010; Haidar et O'Shaughnessy, 2012), la detection des spams (Siefkes *et al.*, 2004; Kolari *et al.*, 2006; Ntoulas *et al.*, 2006; Kanaris *et al.*, 2007; Çıltık et Güngör, 2008; Guzella et Caminhas, 2009; Sohn *et al.*, 2009; Chawla, 2014; Crawford *et al.*, 2015), ainsi que la tâche de datation automatique des textes (Garcia-Fernandez *et al.*, 2011; Kumar *et al.*, 2011; Chambers, 2012; Niculae *et al.*, 2014; Li *et al.*, 2015; Popescu et Strapparava, 2015; Zampieri *et al.*, 2015; Chiru et Toia, 2016; Wahlberg *et al.*, 2016). Ces domaines partagent des similarités, car ils sont axés sur la classification des textes selon divers paramètres comme le style d'auteurs (y compris les bots), le genre, l'époque ou simplement les sujets traités.

Il n'est pas étonnant de retrouver l'analyse des n-grammes dans les domaines de la génétique et de la phylomémétique qui s'intéressent aux aspects liés à l'évolution (respectivement génétique ou pas) pour des tâches comme le séquençage de génomes (Ganapathiraju *et al.*, 2002; Atkinson et Gray, 2005; Volkovich *et al.*, 2005; Tomović *et al.*, 2006; Pagel, 2009; Ding *et al.*, 2011; Howe et Windram, 2011; Chavalarias et Cointet, 2013; Xu *et al.*, 2015).

Parmi les études qui nous intéressent particulièrement, se trouvent celles rattachées au domaine de l'étude diachronique de corpus. Certaines ont pour objectif l'étude de la culture au travers de ces corpus quand d'autres ont une visée plus linguistique. Le travail de recherche (Michel *et al.*, 2011) a donné naissance au domaine Culturomics, l'étude de la culture par l'analyse des profils fréquentiels diachroniques des n-grammes sur le grand corpus de Google Books. Il a fait émerger le potentiel d'inférence d'hypothèses culturelles et linguistiques sur la base de données textuelles du monde réel. Ce travail est complété par l'article (Gao *et al.*, 2012) qui montre qu'une analyse fractale fournit des informations fondamentales sur la nature des corrélations contenues dans les trajectoires du travail à l'origine du domaine Culturomics. Ils concluent que de nouvelles interprétations peuvent être dérivées de la comparaison des trajectoires de ces phénomènes socio-linguistiques vis-à-vis de ceux de la nature.

Parmi les études sur l'évolution sémantique des corpus, le travail (Jatowt et Duh, 2014) effectue une analyse exploratoire afin d'étudier différentes méthodes automatiques d'analyse de l'évolution sémantique des mots et de la langue au niveau lexical. Il propose plusieurs approches améliorant la compréhension de l'évolution diachronique des mots sur de grand corpus de données. Le travail (Kulkarni *et al.*, 2015) propose quant à lui trois approches computationelles de différentes complexités afin de détecter les changements sémantiques des mots, en utilisant diverses caractéristiques de distributions déduites à partir des co-occurrences de mots, sur les n-grammes de divers corpus dont celui de Google Books. Le travail (Hamilton *et al.*, 2016) tente de comprendre la façon dont les mots changent de sens au cours du temps. Il utilise et évalue plusieurs méthodes de type "word embeddings" (méthode d'apprentissage automatique "deep learning" représentant les mots par des vecteurs de nombres réels) afin de révéler des lois statistiques de l'évolution sémantique.

Enfin, d'autres travaux sur les problèmes de représentativité des corpus et la mesure des fréquences (Brysbaert *et al.*, 2011) étudient comment estimer une variables importantes en psychologie expérimentale, la fréquences de mots. L'effet de la sous-représentativité sur le calcul des fréquences est observé ainsi que son lien avec la phase de prétraitement des données. Le travail (Prévost, 2015) traite de l'apport aux études de langues anciennes de la quantification, notamment par la mesure fréquentielle diachronique des mots sur de grand corpus numérisés. Il donne également des indications sur les erreurs d'interprétation potentielles liées à l'étude de ces courbes fréquentielles.

Ces études montrent des développements prometteurs dans l'analyse de l'évolution de la culture et de la linguistique au travers de grands corpus. Toutefois, d'autres pointent les problèmes de représentativité et mettent en garde concernant l'interprétation de ces données.

2.3 Deux variables clés : le niveau n et la taille des corpus

Nous constatons qu'un grand nombre de ces études de corpus diachroniques aborde le niveau *n* des n-grammes comme un paramètre arbitraire. Cette variable joue pourtant un rôle essentiel dans la nature des phénomènes décrits par les n-grammes. Intuitivement, le spectre des niveaux *n* correspond aux niveaux structurants de la langue, c'est-à-dire, lexical, syntaxe, sémantique et pragmatique. Le niveau 1 correspondrait au lexique, les niveaux les plus élevés à la pragmatique tandis que les niveaux intermédiaires correspondraient alors à la syntaxe suivie de la sémantique. Selon cette hypothèse, un niveau *n* intermédiaire peut mettre en évidence des effets de style liés éventuellement à l'auteur ou à la date de création du texte analysé. Le travail (Wijaya et Yeniterzi, 2011) propose de modéliser les changements linguistiques de mots en fonction du changement des mots qui cooccurrent avec celui-ci dans le temps, identifiant des groupes de sujets liés à ces mots sur les n-grammes du corpus de Google Books. Selon ce travail, les 5-grammes permettent de mieux clarifier la nature du changement tout en permettant l'identification plus précise de la période dudit changement. Bien que certains travaux proposent l'utilisation de n-grammes de longueur variable (Siu et Ostendorf, 2000; Marceau, 2001; Jiang *et al.*, 2007; Wang *et al.*, 2013) pour diverses tâches de recherche, ceux-ci se concentrent peu sur la relation existante entre ces différents niveaux. Dans cette thèse, nous étudions donc aussi la relation qu'entretiennent ces niveaux entre eux notamment au travers d'une équation simple reliant les fréquences de certains n-grammes.

Une autre variable décisive dans les études diachroniques de grand corpus est l'évolution de la taille du corpus. Cette variable peut perturber les mesures mises en place et cela peut conduire à constater une évolution perçue comme linguistique ou culturelle, mais qui n'est qu'une conséquence des variations de taille des données généralement plus abondante dans les années les plus récentes. La plupart des études mentionnent à peine le problème ou pire l'ignorent totalement alors que cette variable agit comme une variable cachée et a une influence sur la plupart des mesures utilisées. Dans le travail (Juola, 2013), une mesure de la complexité de la culture et son évolution est étudiée sur les n-grammes du corpus de Google Books. L'hypothèse retenue est que la complexité de la culture augmente avec l'augmentation croissante de la place de la technologie ainsi qu'en raison du rythme de la vie moderne. Toutefois, ce travail ne mentionne pas les effets de taille de corpus existants dans Google Books qui vont influencer l'évolution de l'entropie mesurée. A titre d'exemple, une simple mesure de la diversité lexicale est impactée par l'augmentation de la taille des données constituant l'échantillon de la langue.

Dans cette thèse, nous proposons d'élaborer des méthodes de mesure ayant pour objectif de réduire ou corriger cette sensibilité. Une attention particulière sera employée à ce que les hypothèses présentées ne soient pas le résultat d'effets de variation de taille de corpus, de prétraitement de données ou simplement du bruit. Dans le bruit nous considérons l'évolution des sujets de presse (donnant naissance à une diversité lexicale qui n'est pas un fait linguistique), la variation de la taille du corpus avec le temps, les erreurs d'OCR, l'apparition de nouvelles sections, les annonces officielles redondantes, la publicité, etc.

3 Structure de la thèse

Ce travail de thèse se subdivise en six grande parties. La première est l'introduction. La seconde présente le corpus de presse de 4 millions d'articles sur lequel seront appliqués la plupart des outils, mesures et méthodes développés dans cette thèse. La troisième définit ces mêmes outils et méthodes en explicitant aussi les notions et concepts de base permettant d'appréhender le cadre théorique de la thèse. La quatrième se consacre à l'utilisation de ces méthodes et outils de mesures sur ces corpus en analysant dans un premier temps les mots pour ensuite s'intéresser aux n-grammes de niveau supérieur. La cinquième présente une façon d'agréger ces mesures sur une dimension multi-échelle tout en discutant les relations particulières des n-grammes vis-à-vis de n-grammes d'autres niveaux. Enfin, la dernière est la conclusion.

La structure est donc la suivante :

1. Introduction
2. Corpus
3. Concepts et méthodes
4. Analyse de n-grammes
5. Conclusion et travaux futurs

La deuxième partie commence par un chapitre de présentation générale des données. Nous y présentons le corpus, le prétraitement des données ainsi que diverses études exploratoires menées sur ce corpus. Le second chapitre présente des statistiques extraites sur la base des données textuelles du corpus. Ces diverses statistiques nous permettent une première lecture basique du corpus étudié ainsi que de ses caractéristiques.

La troisième partie présente l'ensemble des notions et concepts que nous utilisons au cours de cette thèse et définit formellement les bases théoriques sur lesquelles s'appuient les nouvelles méthodes développées. Cette partie est également subdivisée en deux chapitres selon ce que nous avons appelé les niveaux Micro et Macro. Ces niveaux se définissent en rapport avec la proximité de l'analyse avec le texte. Le niveau Micro est proche du texte, car il analyse l'évolution fréquentielle des n-grammes et permet de rapidement d'en retrouver de nombreux

exemples d'occurrences au sein de la masse textuelle. Le niveau Macro est éloigné du texte, car les méthodes rattachées considèrent l'ensemble collectif des informations fréquentielles des n-grammes afin de mesurer une évolution globale de la langue sur le corpus donné.

La quatrième partie utilise ces différentes méthodes et outils de mesure sur les corpus de presse selon les niveaux n des n-grammes. La grille de lecture de ce chapitre peut être donnée par un tableau à double entrée dont les lignes représentent le type de méthode utilisée tandis que les colonnes représentent le niveau n des n-grammes. Dans un premier temps l'ensemble de ces méthodes sont appliquées aux mots afin d'analyser l'évolution du lexique. Dans un deuxième temps, chaque méthode sera appliquée successivement sur les n-grammes de niveau supérieur. Ainsi, nous représentons ce tableau à double entrée dans la Figure 3.1.

		Taille des n-grammes		
		1	2-9	multi-échelle
Macro	Distances nucléaires	11.1	12.1	14.1
	Entropie nucléaire	11.2	12.2	
Macro et Micro	Chronocloud	11.3	12.3	14.2
Micro	Profil fréquentiel	11.4	12.4	14.3 et 14.4

FIGURE 3.1 – Tableau de lecture de la partie 4 de la thèse

Cette partie se termine par un chapitre qui tente une réunification des niveau n pour chacune des mesures utilisées. Elle discute également une notion clé : la relation entre les différents niveaux de n-grammes. Cette relation est discutée au travers de deux concepts. Le premier est la décomposition des profils fréquentiels partageant des séquences de mots en commun. La seconde est la notion de corrélation de courbes fréquentielles entre n-grammes ne partageant pas de séquences de mots en commun, c'est-à-dire l'espace complémentaire. Ce chapitre est un point de départ pour d'autres études potentielles sur les n-grammes et leurs relations sur différents niveaux n.

L'objectif des méthodes big data développées dans cette thèse est de mesurer l'évolution linguistique au travers d'une analyse diachronique d'un corpus de presse de 4 millions d'articles s'étalant sur 200 années d'archives. Ainsi, nous précisons que le terme "évolution linguistique" est à prendre au sens de "transformations de la langue" et que nous analysons l'évolution d'un corpus en tentant de cibler autant que possible la langue. De plus, le terme "analyse diachronique" ne fait pas référence à la linguistique classique, mais renvoie simplement à une analyse de l'évolution linguistique d'un corpus s'étalant dans le temps.

4 Contributions

De par les différentes interactions professionnelles effectuées au cours de cette thèse, il est important de mentionner quelques contributeurs indirects à ce travail :

- **Ana Manasovska**, **Fabien Jolidon**, **Florian Junker**, **Joanna Salathé**, **John Gaspoz**, **Mathieu Monney**, **Sidney Bovet**, **Valentin Rutz** et **Zhivka Gucevska** qui ont formé un groupe se consacrant au projet du cours big data sur le sujet "Searching the space of word temporal profiles on Le Temps Newspaper".
- **Ari Sarfatis**, **Christophe Schelling** et **Ciprian Tomoiaga** qui ont formé un groupe se consacrant au projet du cours humanités digitales sur le sujet "Perception de la décolonisation et des nouvelles indépendances dans GDL et JDG durant la Guerre froide".
- **Arnaud Miribel** et **Laurent Valette** qui ont consacré un projet de master sur la mesure de la qualité de l'OCR et l'amélioration de celle-ci sur les corpus de GDL et JDG.
- **Augustin Prado**, **Gina Reuland** et **Jonas Racine** qui ont formé un groupe se consacrant au projet du cours humanités digitales sur le sujet "Evolution linguistique du débat public autour de l'immigration en Suisse".
- **Ciprian Tomoiaga** qui a consacré un projet de semestre sur la notion de chronocloud et de noyau résilient appliqué au corpus de Google Books multilingue.
- **Cynthia Oeschger**, **Farah Bouassida**, **Gil Brechbühler**, **Jérémy Weber**, **Malik Bougacha**, **Marc Schär**, **Nicolas Bornand** et **Tao Lin** qui ont formé un groupe se consacrant au projet du cours big data sur le sujet "A Study of Linguistic Drift On Le Temps Newspaper Corpus".
- **Cyril Bornet** pour les réflexions à propos de la notion de noyau résilient et de chronocloud.
- **Dario Rodighiero** pour les réflexions sur le design du n-grammes viewer et du chronocloud.
- **Maud Ehrmann**, **Giovanni Colavizza** et **Yannick Rochat** pour les nombreux conseils.
- **Patricia Marlet** pour les corrections orthographiques et de grammaire de cette thèse.
- **Patrick Staeger** pour l'aide sur l'aspect "gestion de projet" et les corrections orthographiques et de grammaire de cette thèse.
- **Pierre Runavot** pour les nombreuses réflexions à propos de l'étude de l'évolution linguistique sur les corpus de GDL et JDG.

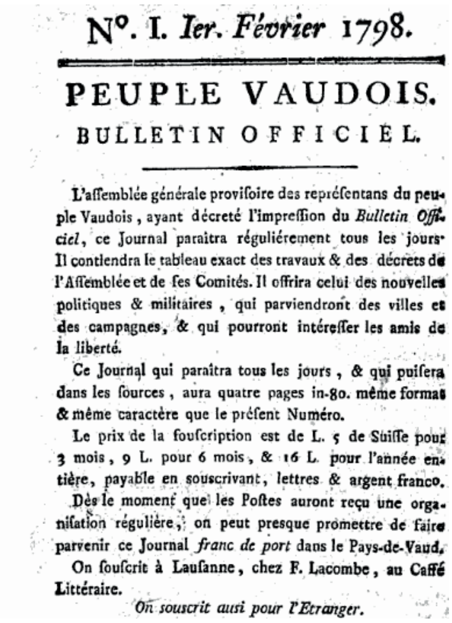
Corpus Partie II

5 Introduction aux données

Dans ce chapitre, nous introduisons les données relatives aux deux corpus de presse que nous utilisons pour la plupart des analyses effectuées dans cette thèse afin de mieux comprendre la masse d'informations contenue dans ceux-ci. Nous commençons par présenter les deux journaux et l'historique du projet presse. Nous présenterons ensuite une méthode d'étude de l'évolution de la mise en page des premières pages de ces journaux permettant d'analyser la structure du média indépendamment de son contenu. Enfin, nous présenterons les données textuelles des corpus ainsi que les prétraitements que nous avons appliqués à ces données.

5.1 Présentation des corpus

Nous utilisons un corpus composé de 4 millions d'articles de presse documentant indirectement l'évolution de la langue française écrite sur 200 années de 1798 à 1998. Le corpus est composé de deux journaux, la "Gazette de Lausanne" (GDL) allant de 1798 à 1998 et le "Journal de Genève" (JDG) allant de 1826 à 1998. L'archive est composée de pages numérisées en format PDF et PNG ainsi que de fichiers XML contenant leurs données textuelles. Ces données sont le résultat d'un processus de reconnaissance optique de caractères (OCR / Optical Character Recognition) ainsi que d'algorithmes de détection de mise en page et d'un prétraitement de tokenisation (détection des unités lexicales). Ces fichiers XML fournissent la position des articles et des informations structurelles tandis que d'autres fichiers XML fournissent pour chaque article tous les tokens (unités lexicales), leurs positions et diverses métadonnées (numéro de page, date de parution, résolution d'image, nombre de mots, nombre de caractères, nombre de caractères potentiellement erronés, etc). Les articles commençant sur une page du journal et se terminant sur une autre sont reconstitués en un unique article. Les résultats du processus d'OCR, la détection de la mise en page et la tokenisation ont été fournies par le journal Le Temps, propriétaire des données. Ces archives représentent aussi 100 000 parutions, 1 million de pages, 2.5 milliards de mots, 440 000 images, 17 gigaoctets de texte et au total 25 téraoctets de données. Les données textuelles constituent la base principale de la majorité des analyses effectuées au cours de cette thèse. Les premières pages des archives des deux journaux sont présentées dans les Figures 5.1 et 5.2.



N°. I. *Ier. Février* 1798.

PEUPLE VAUDOIS.

BULLETIN OFFICIEL.

L'aſſemblée générale proviſoire des repréſentans du peuple Vaudois, ayant décrété l'impreſſion du *Bulletin Officiel*, ce Journal paraitra réguliérement tous les jours. Il contiendra le tableau exact des travaux & des décrets de l'Aſſemblée et de ſes Comités. Il offrira celui des nouvelles politiques & militaires, qui parviendront des villes et des campagnes, & qui pourront intéreſſer les amis de la liberté.

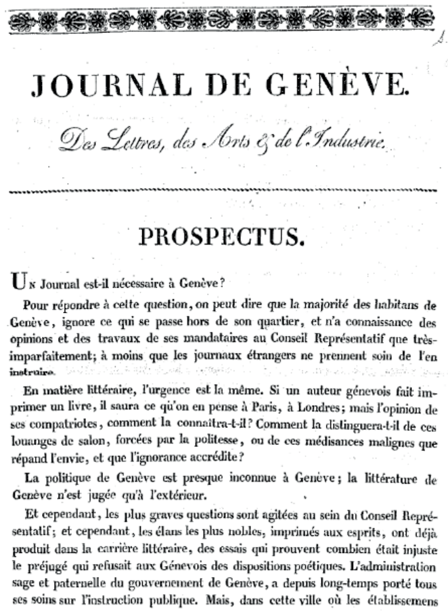
Ce Journal qui paraîtra tous les jours, & qui puiſera dans les ſources, aura quatre pages in-8o. même format & même caractère que le préſent Numéro.

Le prix de la ſouſcription est de L. 5 de Suiſſe pour 3 mois, 9 L. pour 6 mois, & 16 L. pour l'année entière, payable en souscrivant, lettres & argent franco.

Dès le moment que les Poſtes auront reçu une organiſation régulière, on peut presque promettre de faire parvenir ce Journal *franc de port* dans le Pays-de-Vaud.

On ſouſcrit à Lauſanne, chez F. Lacombe, au Caffé Littéraire.

On souscrit auſsi pour l'Etranger.



JOURNAL DE GENÈVE.

Des Lettres, des Arts & de l'Industrie.

PROSPECTUS.

Un Journal est-il nécessaire à Genève?

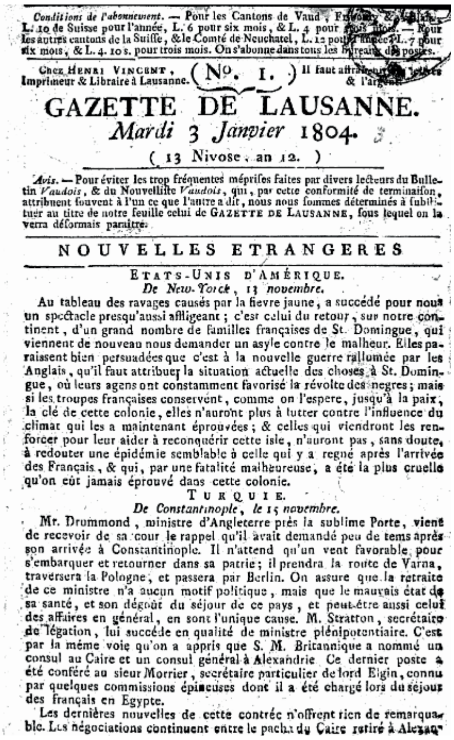
Pour répondre à cette question, on peut dire que la majorité des habitans de Genève, ignore ce qui se passe hors de son quartier, et n'a connaissance des opinions et des travaux de ses mandataires au Conseil Représentatif que très-imparfaitement; à moins que les journaux étrangers ne prennent soin de l'en instruire.

En matière littéraire, l'urgence est la même. Si un auteur génevois fait imprimer un livre, il saura ce qu'on en pense à Paris, à Londres; mais l'opinion de ses compatriotes, comment la connaîtra-t-il? Comment la distinguera-t-il de ces louanges de salon, forcées par la politesse, ou de ces médisances malignes que répand l'envie, et que l'ignorance accrédite?

La politique de Genève est presque inconnue à Genève; la littérature de Genève n'est jugée qu'à l'extérieur.

Et cependant, les plus graves questions sont agitées au sein du Conseil Représentatif; et cependant, les élans les plus nobles, imprimés aux esprits, ont déjà produit dans la carrière littéraire, des essais qui prouvent combien était injuste le préjugé qui refusait aux Génevois des dispositions poétiques. L'administration sage et paternelle du gouvernement de Genève, a depuis long-temps porté tous ses soins sur l'instruction publique. Mais, dans cette ville où les établissemens

FIGURE 5.1 – Premières pages des archives de GDL (gauche) le 1er février 1798 et de JDG (droite) le 1er janvier 1826



Conditions de l'abonnement. — Pour les Cantons de Vaud, Fribourg & ..., L. 10 de Suisse pour l'année, L. 6 pour six mois, & L. 4 pour trois mois. — Pour les autres cantons de la Suiſſe, & le Comté de Neuchatel, L. 12 pour l'année, L. 7 pour six mois, & L. 4. 10 s. pour trois mois. On s'abonne dans tous les bureaux des postes.

Chez Henri Vincent, Imprimeur & Libraire à Lausanne. (N°. 1.) Il faut affranchir les lettres & l'argent.

GAZETTE DE LAUSANNE.

Mardi 3 *Janvier* 1804.

(13 Nivose, an 12.)

Avis. — Pour éviter les trop fréquentes mépriſes faites par divers lecteurs du Bulletin *Vaudois*, & du Nouvelliſte *Vaudois*, qui, par cette conformité de terminaiſon, attribuent ſouvent à l'un ce que l'autre a dit, nous nous ſommes déterminés à ſubſtituer au titre de notre feuille celui de GAZETTE DE LAUSANNE, ſous lequel on la verra déſormais paraître.

NOUVELLES ETRANGERES

ETATS-UNIS D'AMÉRIQUE.

De New-Yorck, 13 novembre.

Au tableau des ravages causés par la fievre jaune, a succédé pour nous un spectacle presqu'aussi affligeant; c'est celui du retour, sur notre continent, d'un grand nombre de familles françaises de St. Domingue, qui viennent de nouveau nous demander un asyle contre le malheur. Elles paraissent bien persuadées que c'est à la nouvelle guerre rallumée par les Anglais, qu'il faut attribuer la situation actuelle des choses à St. Domingue, où leurs agens ont constamment favorisé la révolte des negres; mais si les troupes françaises conservent, comme on l'espere, jusqu'à la paix, la clé de cette colonie, elles n'auront plus à lutter contre l'influence du climat qui les a maintenant éprouvées; & celles qui viendront les renforcer pour leur aider à reconquérir cette isle, n'auront pas, sans doute, à redouter une épidémie semblable à celle qui y a regné après l'arrivée des Français, & qui, par une fatalité malheureuse, a été la plus cruelle qu'on eût jamais éprouvé dans cette colonie.

TURQUIE.

De Constantinople, le 15 novembre.

Mr. Drummond, ministre d'Angleterre près la sublime Porte, vient de recevoir de sa cour le rappel qu'il avait demandé peu de tems après son arrivée à Constantinople. Il n'attend qu'un vent favorable pour s'embarquer et retourner dans sa patrie; il prendra la route de Varna, traversera la Pologne, et passera par Berlin. On assure que la retraite de ce ministre n'a aucun motif politique, mais que le mauvais état de sa santé, et son dégoût du séjour de ce pays, et peut-être aussi celui des affaires en général, en sont l'unique cause. M. Stratton, secrétaire de légation, lui succède en qualité de ministre plénipotentiaire. C'est par la même voie qu'on a appris que S. M. Britannique a nommé un consul au Caire et un consul général à Alexandrie. Ce dernier poste a été conféré au sieur Morrier, secrétaire particulier de lord Elgin, connu par quelques commissions épineuses dont il a été chargé lors du séjour des français en Egypte.

Les dernières nouvelles de cette contrée n'offrent rien de remarquable. Les négociations continuent entre le pacha du Caire retiré à Alexan-

JEUDI, 5 Janvier 1826.

Ire ANNÉE. N.° 1.

JOURNAL DE GENEVE

DES LETTRES, DES ARTS ET DE L'INDUSTRIE.

ÉPHÉMÉRIDES. 1791 Janvier 5. On a proposé en conseil des Deux Cents, d'abolir le décret rendu contre Jean-Jacques Rousseau, et de lui élever une statue sur le pied de laquelle on inscrira qu'elle est destinée à effacer l'outrage qu'il a reçu de cet État. (Genève.)

CONSEIL REPRÉSENTATIF.

Session ordinaire de décembre 1825.

SÉANCE DU 7 DÉCEMBRE.

SÉANCE DU 9 DÉCEMBRE.

SÉANCE DU 12 DÉCEMBRE.

FIGURE 5.2 – Premières pages du lancement définitif de GDL (gauche) le 3 janvier 1804 et de JDG (droite) le 5 janvier 1826

Nous observons que les toutes premières pages sont rudimentaires au niveau de la mise en page ainsi qu'au niveau de la police utilisée notamment par GDL rendant le caractère "s" difficile à reconnaître par le processus d'OCR, car il ressemble plus au caractère "f" même pour un lecteur humain. Nous observons également que les premières données de JDG ne sont pas le journal lui-même, mais un dépliant annonçant le lancement du journal qui est finalement paru quatre jours plus tard, le 5 janvier 1926. Pour GDL, le journal avait d'abord le nom de "Bulletin officiel", ensuite "Journal helvétique" avant de prendre finalement le nom de "Gazette de Lausanne" le 3 janvier 1804. Les deux journaux ont fusionné en 1991 et sont devenus le "Journal de Genève et Gazette de Lausanne", conservant néanmoins quelques différences au niveau d'articles traitant de sujets locaux. Ils deviennent définitivement un seul et même journal en 1998 prenant le nom du journal "Le Temps". Une illustration des unes de GDL sur diverses années est présentée dans la Figure 5.3.

FIGURE 5.3 – Unes de GDL en 1825 (haut / gauche), 1850 (haut / milieu), 1875 (haut / droite), 1925 (bas / gauche), 1950 (bas / milieu) et 1975 (bas / droite)

Nous constatons que les changements de la mise en page des premières pages de GDL sont importants sur ces parutions séparées par 25 années. Nous observons en particulier des changements de police et de taille de caractères tant au niveau du texte que du titre. Nous observons aussi un nombre de colonnes évoluant avec le temps impliquant potentiellement des variations de qualité de l'OCR. En règle générale, l'évolution touche également le nombre d'articles et de pages, la longueur des articles et les ajouts de nouvelles sections, de documents supplémentaires, d'éditions spéciales et littéraires, etc.

Les corpus de JDG et GDL sont mis à la disposition de la recherche dans le cadre du projet presse né de la collaboration entre la Bibliothèque Nationale Suisse (BNS), le journal Le Temps et le laboratoire des humanités digitales de l'EPFL (DHLAB). L'objectif de ce projet est de transformer ces 4 millions d'articles en un système d'information.

En effet, à la suite de les fusion des deux journaux JDG et GDL. Le journal Le Temps et ses partenaires ont fait naître le projet presse selon la chronologie suivante :
- 2005 : étude sur la numérisation par la BNS
- 2006 : numérisation du "Journal de Genève"
- 2008-2009 : numérisation de la "Gazette de Lausanne" et du "Nouveau Quotidien"
- 2011 : l'EPFL participe à la rédaction d'un projet national sur ces corpus dans le domaine des humanités digitales
- 2012 : convention de recherche entre EPFL et Le Temps
- 2013 : création du comité scientifique de Le Temps
- 2014 : financement de la BNS sur le projet

Cette thèse est financée par le Fonds National Suisse (FNS) dans le cadre du projet "How algorithms shape langage", numéro 149758. Ce projet a pour objectif de mesurer les changements linguistiques récents afin de déterminer comment les algorithmes, intervenant de manière croissante sur notre expression textuelle, changent notre langage naturel.

Les archives du projet presse constituent une ressource particulièrement intéressante pour la recherche, pouvant être exploitée pour diverses études dont certaines ont déjà été réalisées afin de mettre en lumière des questions de recherches variées et notamment celles concernant l'analyse linguistique de corpus.

Des articles de conférence ont été publiés sur l'analyse de n-grammes (Buntinx et Kaplan, 2015) et sur les concepts de noyau résilient et la résilience des mots (Buntinx *et al.*, 2016). Un article de journal scientifique a également été publié sur ce dernier concept de noyau et mots résilients (Buntinx *et al.*, 2017a). Cette thèse s'appuie sur ces articles afin de développer la partie théorique et méthodologique.

D'autres études portent sur la conservation d'archives numériques (Rochat *et al.*, 2016) et la reconnaissance des entités nommées (Ehrmann *et al.*, 2016). Enfin, une étude a été publiée à propos de la détection automatique de l'évolution de la mise en page (Buntinx *et al.*, 2017b) et est présentée dans la prochaine section.

5.2 Evolution de la mise en page

Les corpus textuels de GDL et JDG sont sujets à la fiabilité des résultats de la transcription effectuée par la reconnaissance optique de caractères (OCR). Comme nous étudions l'évolution linguistique du corpus, il est légitime de se poser des questions quant à la variabilité de ces résultats au cours du temps. En outre, des facteurs comme la taille de la police ainsi que la qualité et la résolution des parutions scannées peuvent être liés à la qualité de l'OCR. Certains de ces facteurs sont dépendants de l'évolution de la mise en page du journal. En parallèle à l'étude de l'évolution linguistique du corpus, nous avons élaboré une méthode mathématiquement simple permettant d'analyser l'évolution de la mise en page des deux journaux. Cette méthode a également été présentée dans l'article (Buntinx *et al.*, 2017b).

Cette section a donc pour objectif de développer une méthode de détection automatique de la mise en page, nous permettant d'étudier la structure d'un média indépendamment de son contenu alors que dans cette thèse nous étudions plutôt l'évolution de son contenu indépendamment de sa structure. En effet, notre approche se base sur l'image bitmap scannée du journal et n'est pas affectée par l'OCR. Nous allons ainsi tenter de reconstruire une partie de la stratégie éditoriale des journaux et retracer les événements majeurs de leur histoire en terme de changement de mise en page. Nous utiliserons la même méthodologie afin de comparer l'évolution diachronique des deux corpus en temps réel.

La Figure 5.3 permet de visualiser en quelques exemples de premières pages à quel point la mise en page peut changer en 25 années d'écart. Dans certains cas, la transformation est radicale. Nous observons rapidement qu'il est principalement question de position et police du titre, de police de caractères, d'emplacement de la table des matières, de publicités ou de sous-titres et du nombre de colonnes. L'idée est d'utiliser une approche d'analyse factorielle de bitmap. Nous déterminons, dans une première étape, une représentation statistique des images des premières pages en effectuant la moyenne des pixels. Cela signifie que chaque pixel de l'image représentant par exemple un mois est calculé comme la moyenne des pixels exactement situés au même endroit, mais selon les différents jours du mois en question. Le même raisonnement s'applique à la représentation moyenne annuelle. Nous pouvons exprimer la représentation moyenne mensuelle à l'aide de l'équation suivante :

$$\overline{P^t_{y,m}} = \frac{1}{N_{y,m}} \sum_{d=1}^{N_{y,m}} P^t_{y,m,d}$$

De façon similaire, la représentation moyenne annuelle s'exprime par l'équation suivante :

$$\overline{\overline{P^t_y}} = \frac{1}{N_y} \sum_{m=1}^{N_y} \overline{P^t_{y,m}}$$

Une illustration du processus de création des représentations moyennes est représenté dans la Figure 5.4 ainsi que l'image moyenne annuelle de GDL pour trois années dans la Figure 5.5.

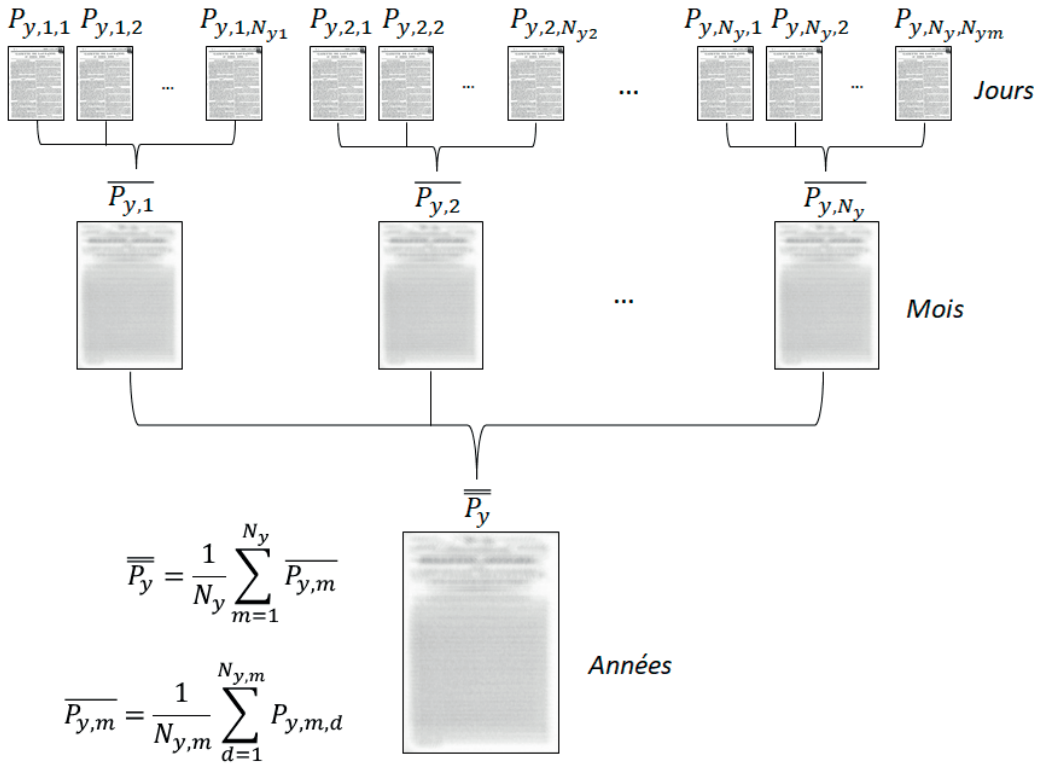


FIGURE 5.4 – Processus de construction d'une représentation annuelle à l'aide des premières pages de chaque parution journalière du journal considéré

FIGURE 5.5 – Représentation moyenne annuelle de GDL pour l'année 1948 (gauche), 1949 (milieu) et 1950 (droite)

Nous observons par exemple l'évolution de la représentation moyenne annuelle de GDL passant de 6 colonnes en 1948 à 7 colonnes en 1950. Nous remarquons aussi que le coin inférieur gauche devient plus clair en 1950. Cet effet est constaté à cet emplacement, car la table des matière du journal est presque toujours au même endroit dans le coin inférieur gauche. Le titre qui possède une certaine inertie dans le temps est lisible sur ces représentations statistiques, mais le texte est totalement flou ce qui permet de faire abstraction du contenu et d'analyser uniquement l'évolution de la structure.

La granularité temporelle reste un choix arbitraire si ce n'est que l'avantage d'une représentation statistique annuelle est sa stabilité et elle permet de repérer des changements de mise en page sur le long terme en évitant trop de variations statistiques. De plus, nous travaillons également dans cette thèse à une échelle annuelle pour déterminer l'évolution fréquentielle des mots et n-grammes, il est donc pertinent d'analyser l'évolution de la structure des journaux JDG et GDL également sur une échelle annuelle.

La seconde étape de la méthode est directement inspirée du principe des "eigenfaces" (visages propres) utilisé pour la reconnaissance automatique des visages (Turk et Pentland, 1991a)(Turk et Pentland, 1991b). Nous calculons les vecteurs propres et les valeurs propres associés à la matrice de la covariance des pixels provenant des représentations annuelles. Les vecteurs propres, par analogie aux eigenfaces, sont appelé eigenpages (pages propres).

Nous projetons ensuite cet espace par analyse en composantes principales sur un plan de dimension 2. Afin de maximiser la covariance des pixels la projection est effectuée selon les deux vecteurs propres qui ont les valeurs propres les plus élevées. Les deux pages propres utilisés pour la projection de JDG et GDL sont représentés dans la Figure 5.6.

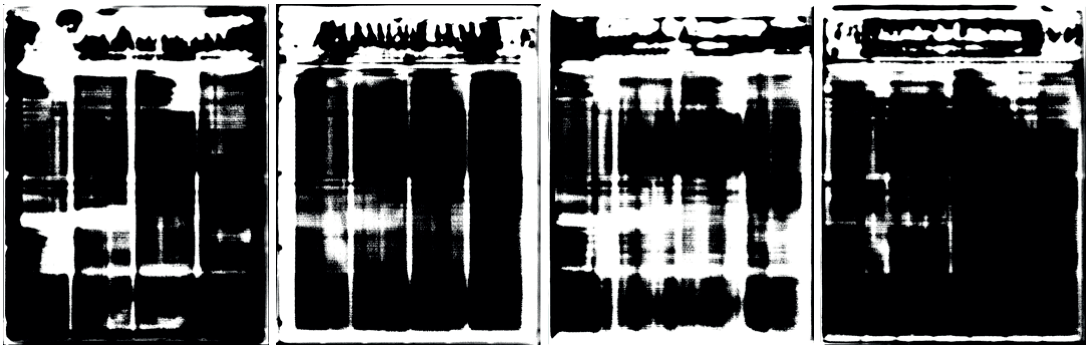


FIGURE 5.6 – Pages propres dont les valeurs propres sont les plus élevées pour le journal de JDG (gauche) et celui de GDL(droite)

La dernière étape consiste à repérer les groupes (clusters) d'années qui sont proches et donc considérées dans une même catégorie structurelle. Cela peut se faire par le biais de la visualisation du plan de la projection. Le plan de projection des journaux de JDG et GDL sur 99 années de 1900 à 1998 ainsi que les regroupements identifiés par inspection manuelle sont présentés respectivement dans les Figures 5.7 et 5.8.

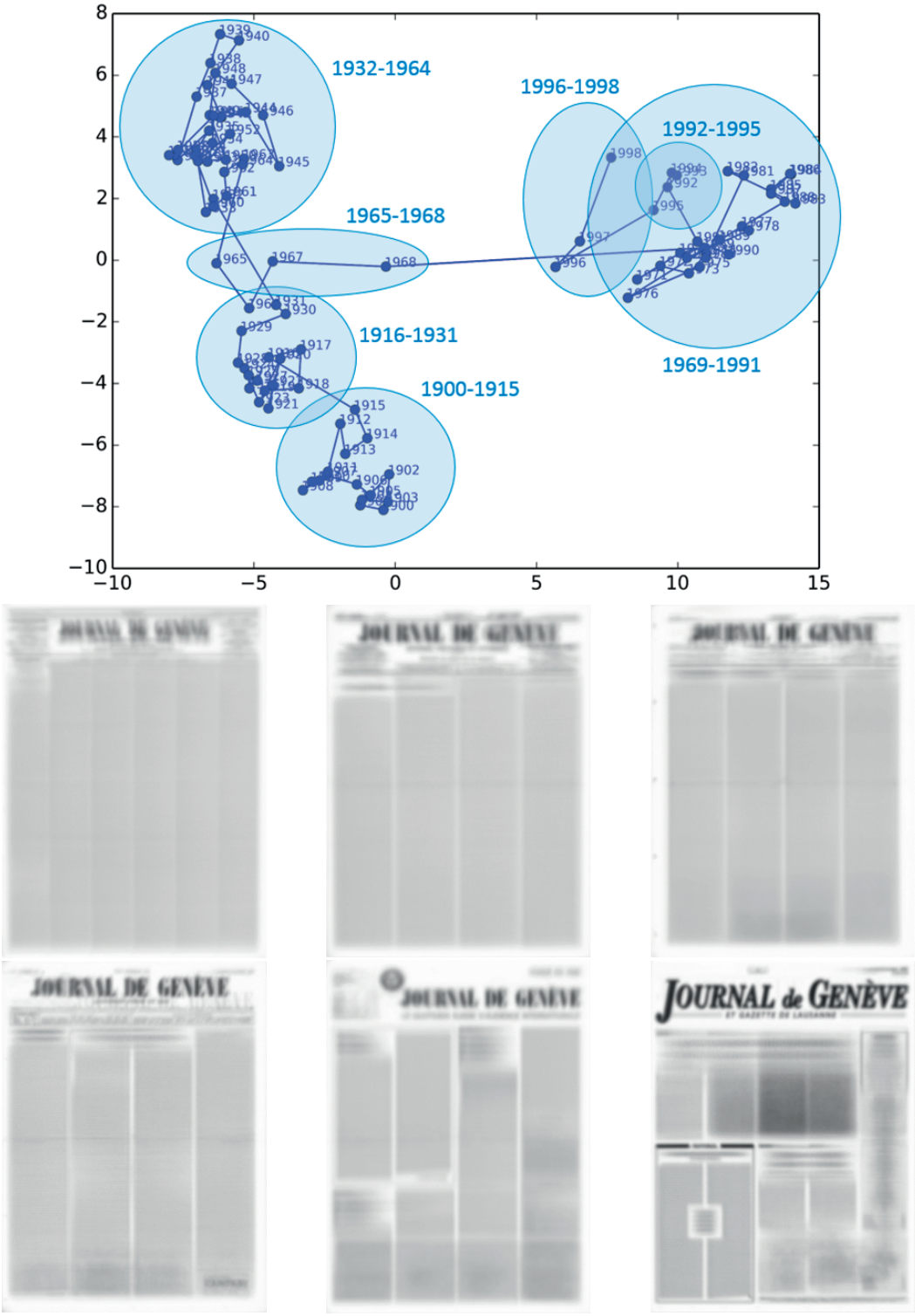


FIGURE 5.7 – Projection des années 1900 à 1998 du journal de JDG (information : 73%) ainsi que les représentations statistiques correspondant aux 6 premiers groupes (clusters) principaux identifiés manuellement, 1900-1915 (haut / gauche), 1916-1931 (haut / milieu), 1932-1964 (haut / droite), 1965-1968 (bas / gauche), 1969-1991 (bas / milieu) et 1992-1995 (bas / droite)

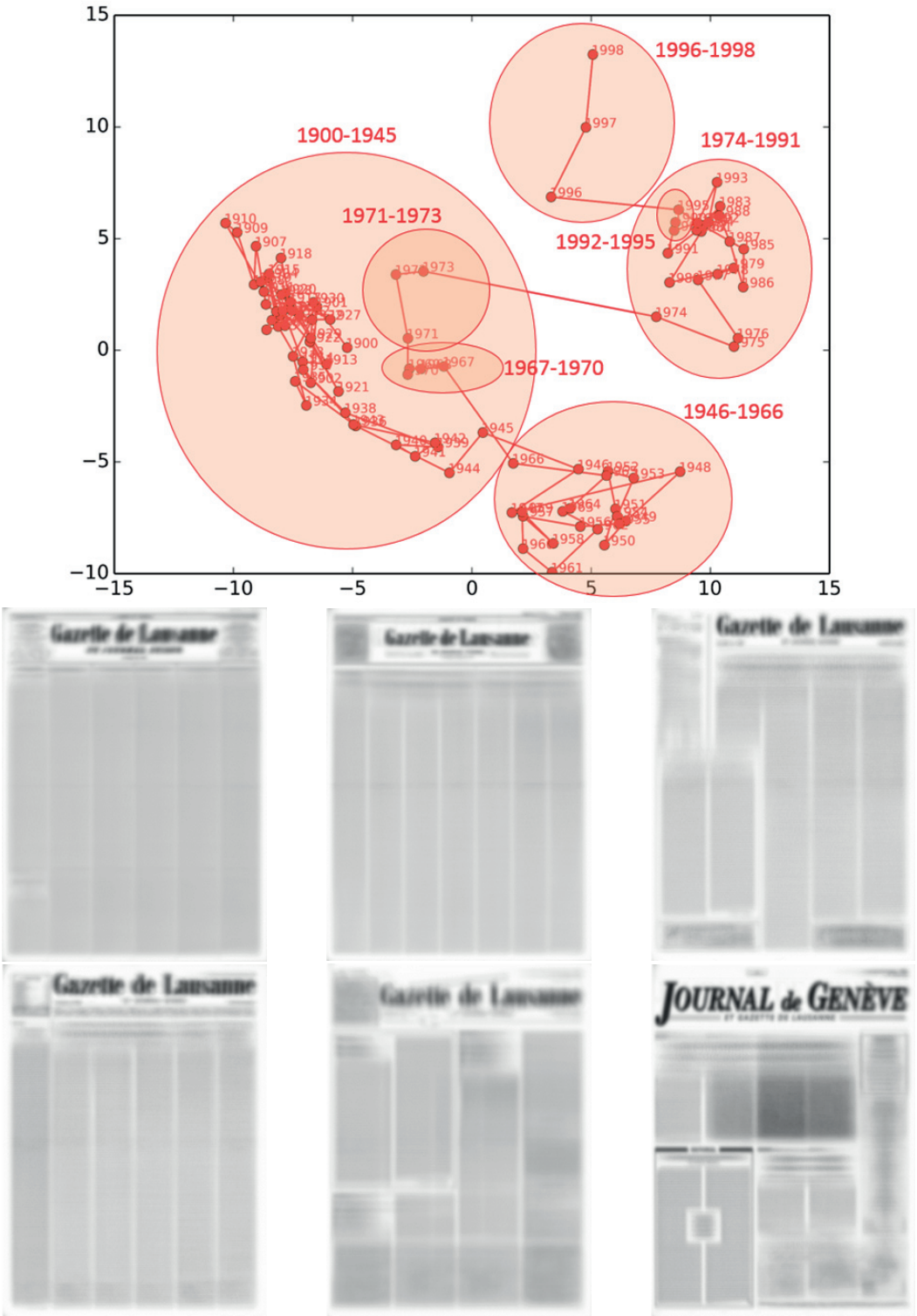


FIGURE 5.8 – Projection des années 1900 à 1998 du journal de GDL (information : 76%) ainsi que les représentations statistiques correspondant aux 6 premiers groupes (clusters) principaux identifiés manuellement, 1900-1945 (haut / gauche), 1946-1966 (haut / milieu), 1967-1970 (haut / droite), 1971-1973 (bas / gauche), 1974-1991 (bas / milieu) et 1992-1995 (bas / droite)

Dans la projection de JDG, nous observons divers groupes d'années consécutives qui sont proches, car elles ont une mise en page similaire. Une regroupement adéquat de ces années permet d'observer les périodes durant lesquelles la mise en page reste semblable ainsi que les transitions qui opèrent entre différents types de mise en page. La distance entre ces points permet de mesurer l'ampleur du changement de mise en page lors de ces périodes de transition. Nous identifions environ 7 groupes d'années dont les deux derniers sont fortement similaires pour les deux journaux, car ceux-ci ont fusionné en 1991. En analysant ces groupes, nous remarquons que la position et police des titres ainsi que le nombre de colonnes constituent des caractéristiques dont les changements ont des effets importants sur la position de la représentation annuelle dans la projection. Nous notons notamment que GDL a longtemps conservé un nombre important de colonnes alors que JDG a rapidement privilégié la lisibilité en les diminuant à 4. GDL a aussi montré qu'il pouvait essayer un format différent avant de revenir en arrière. Il est intéressant de faire un lien entre ces différentes tendances et la réputation des deux journaux, JDG étant considéré comme plus conservateur que GDL. Les groupes identifiés ainsi que leurs caractéristiques sont les suivants :

Journal de Genève (JDG) :
- 1900-1915 : 6 colonnes, titre au dessus des colonnes 2 à 5, peu d'espace entre les colonnes.
- 1916-1931 : 4 colonnes, titre au dessus des colonnes 1 à 4, plus d'espace entre les colonnes.
- 1932-1964 : 4 colonnes, changement de la mise en page autour du titre et de la position du premier sous-titre.
- 1965-1968 : 4 colonnes, changement de la mise en page autour du titre, un cadre avec des bordures noires commence à apparaître.
- 1969-1991 : 4 colonnes, changement de la mise en page du titre, titre au dessus des colonnes 2 à 4, un logo apparaît, davantage d'espace entre les colonnes et les cadres, les titres des articles sont plus grands.
- 1992-1995 : 5 colonnes, fusion de JDG et GDL, refonte totale de la mise en page.
- 1996-1998 : 6 colonnes, changement dans la police du titre, mise en page plus classique au niveau des colonnes, les titres d'articles sont placés dans le haut de la page.

Gazette de Lausanne (GDL) :
- 1900-1945 : 6 colonnes, titre au-dessus des colonnes 2 à 5, peu d'espace entre les colonnes.
- 1946-1966 : 7 colonnes, titre au-dessus des colonnes 2 à 6, plus d'espace entre colonnes donnant des tailles de colonnes particulièrement petites.
- 1967-1970 : 5 colonnes, titre au-dessus des colonnes 2 à 5, la première colonne commence avant le titre qui se trouve à droite, publicités placées en bas de la page.
- 1971-1973 : 6 colonnes, mise en page plus classique avec les titres d'articles en haut de page.
- 1974-1991 : 4 colonnes, beaucoup d'espace entre les colonnes et les articles, les titres d'articles sont plus grands.
- 1992-1995 : 5 colonnes, fusion de JDG et GDL, refonte totale de la mise en page.
- 1996-1998 : 6 colonnes, changement dans la police du titre, mise en page plus classique au niveau des colonnes, les titres d'articles son placés dans le haut de la page.

Afin de comparer les deux journaux plus efficacement nous avons projeté ceux-ci dans un seul et même plan. Cela a pour conséquence de permettre l'interprétation directe des distances entre les représentations annuelles des deux journaux. Toutefois, une telle projection comporte deux fois plus de points et l'information conservée par la projection en est donc diminuée. Le plan de projection des journaux de JDG et GDL sur 99 années de 1900 à 1998 dans un plan unique est présenté dans la Figure 5.9.

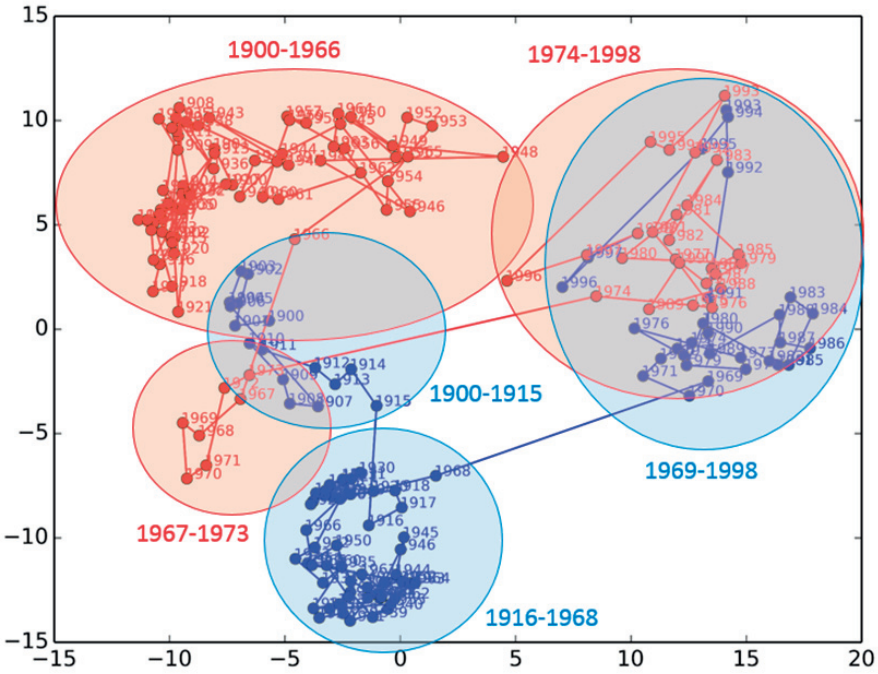


FIGURE 5.9 – Projection des années 1900 à 1998 de JDG et GDL dans un plan unique (information : 67%)

Vu le nombre de points projetés et l'information disponible, seuls trois grands groupes pour chaque journal ont pu être différenciés sur le plan de projection. Ces groupes ont toutefois un certain décalage en terme de temporalité et seul le dernier groupe se confond entre les deux journaux. En effet, le dernier groupe correspond à une mise en page extrêmement proche dès 1974, alors que les deux journaux ne fusionnent qu'en 1991. Nous observons également que c'est le journal de GDL qui s'est aligné sur la mise en page du journal de JDG en 1974 alors que JDG avait déjà adopté cette mise en page dès 1969.

Afin de comparer l'évolution "en temps réel" des deux journaux, nous avons créé une vidéo montrant le plan commun de projection ainsi que les représentations annuelles pour chaque année (avec une mémoire des années passées pour le plan de projection). Les parties correspondant aux années 1964 à 1975 sont présentées dans les Figures 5.10, 5.11 et 5.12.

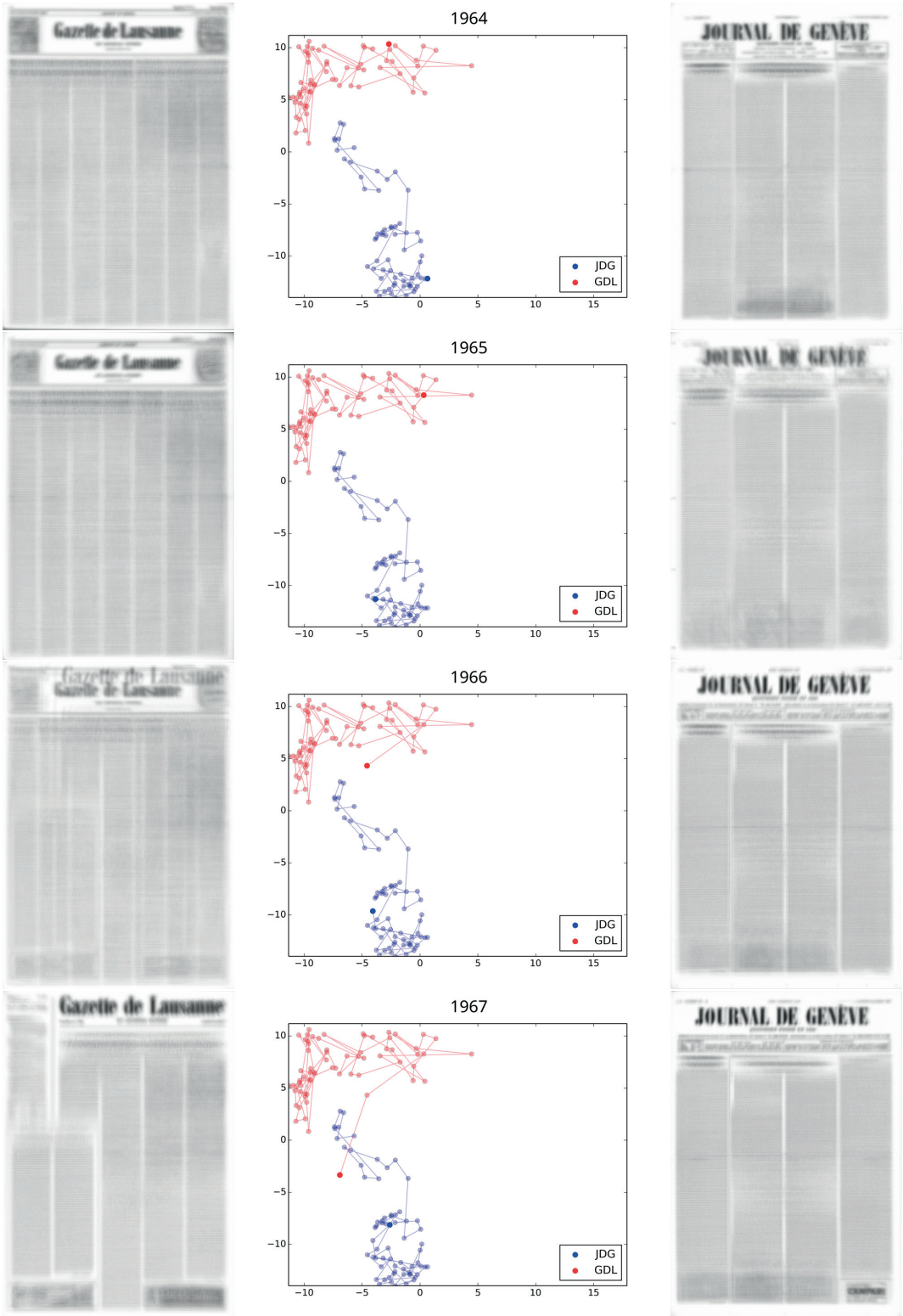


FIGURE 5.10 – Plan commun de projection de JDG et GDL ainsi que les représentations annuelles pour chaque année de 1964 à 1967

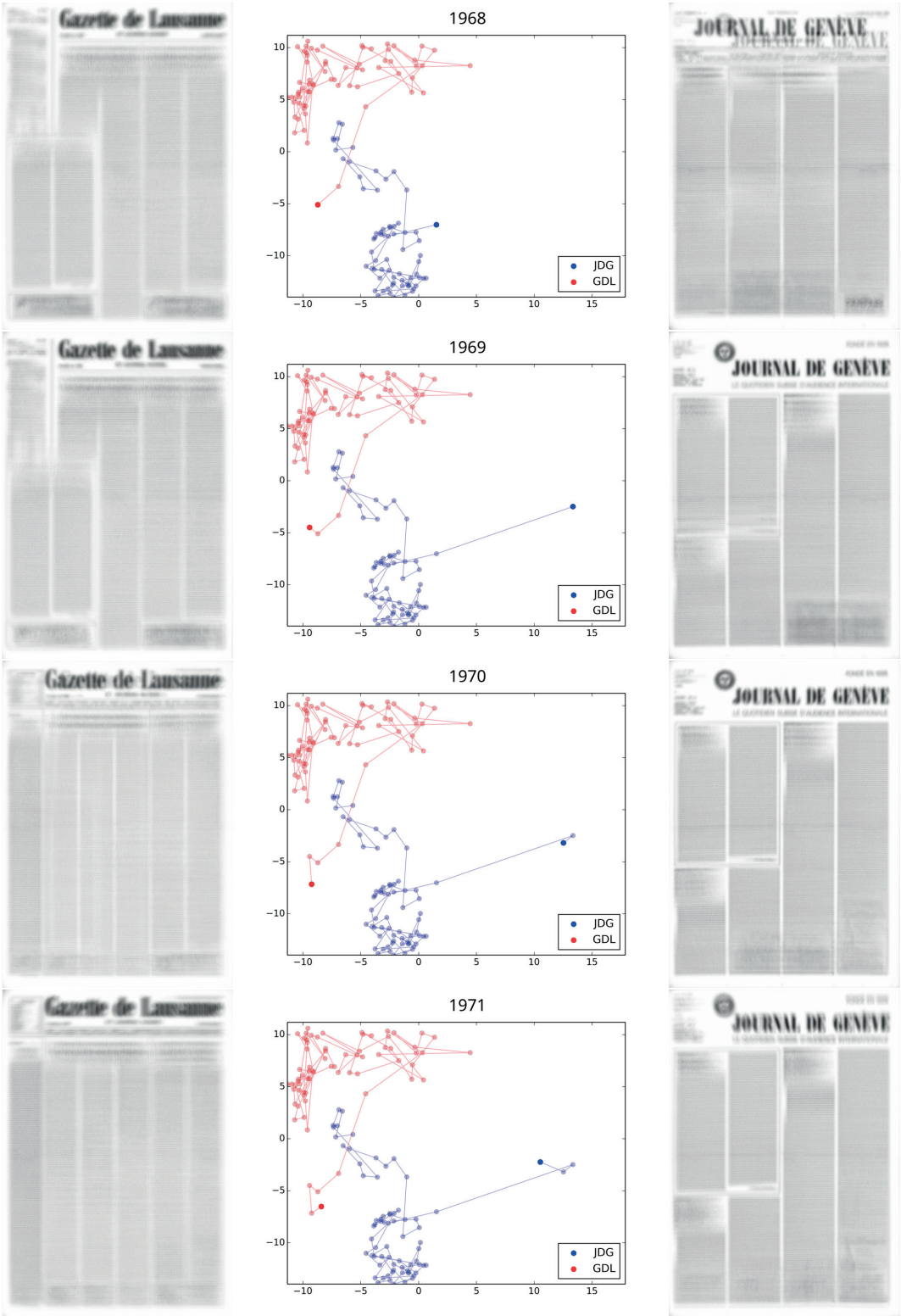


FIGURE 5.11 – Plan commun de projection de JDG et GDL ainsi que les représentations annuelles pour chaque année de 1968 à 1971

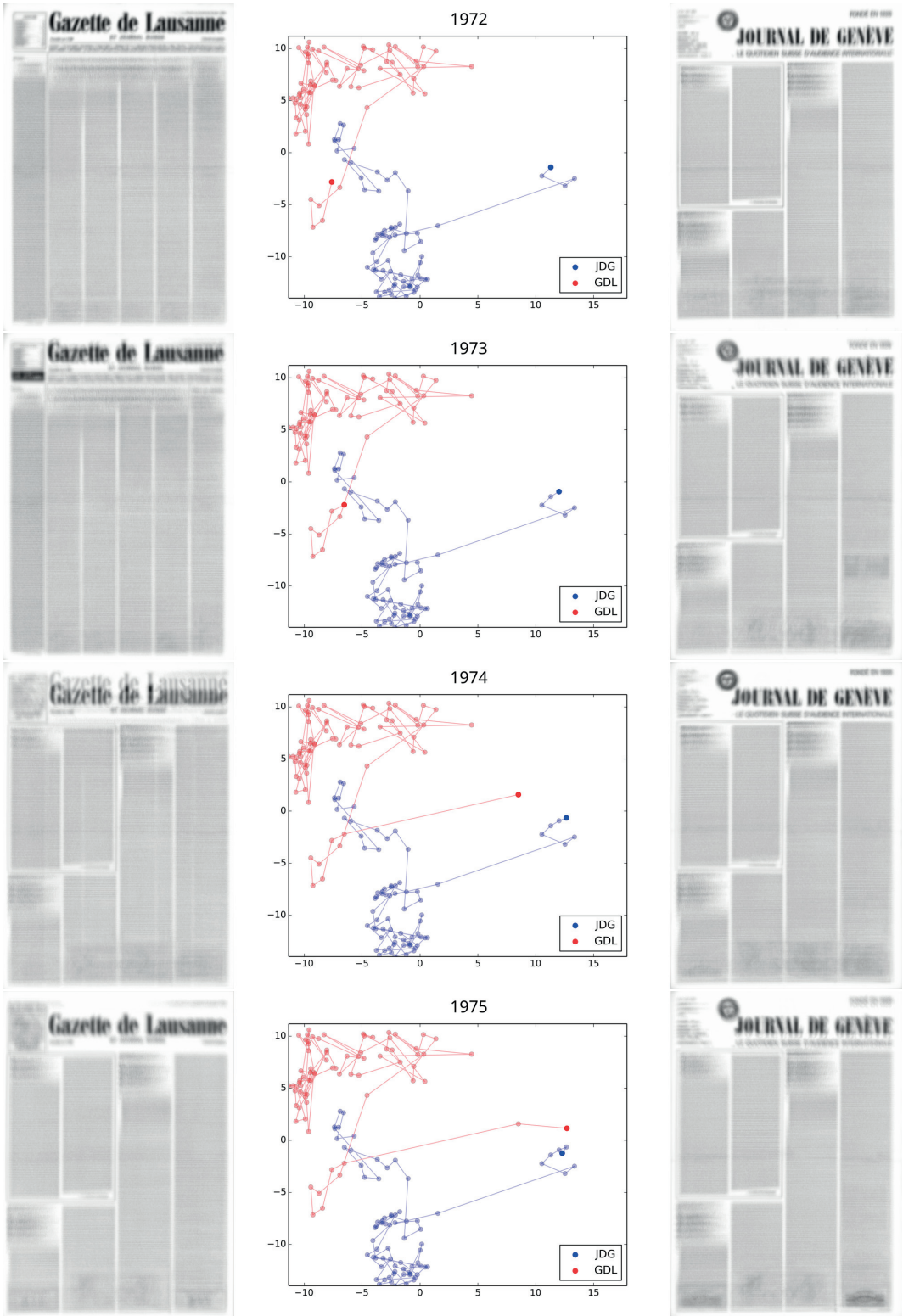


FIGURE 5.12 – Plan commun de projection de JDG et GDL ainsi que les représentations annuelles pour chaque année de 1972 à 1975

La méthode permet de détecter des transitions de mise en page importantes et se révèle un succès dans la comparaison des stratégies éditoriales des deux journaux en terme de mise en page. Elle nous a permis de retracer une partie de l'histoire des journaux indépendamment du contenu des journaux. Ces résultats sont donc aussi totalement indépendants du processus d'OCR lors de notre analyse de l'évolution du corpus basé sur son contenu. La projection des deux journaux sur un même plan et la vidéo retraçant en temps réel les évolutions de mise en page nous ont permis d'approfondir la relation qu'entretiennent ces deux journaux dans un espace-temps proche (Genève et Lausanne sont séparés d'environ 60 km). En outre, cette méthode a l'avantage d'être générique et mathématiquement simple, ce qui permet de l'appliquer à d'autres corpus de journaux afin de les comparer. Il est également possible d'utiliser les meta-données dont nous disposons sur les positions des images afin de construire une représentation annuelle ne tenant compte que de la composante "images" ce qui ouvre la possibilité de décomposer l'évolution de la mise en page selon différents axes comme celui de l'emplacement des images. Un exemple de ces représentations est illustré dans la Figure 5.13.

FIGURE 5.13 – Représentation annuelle de GDL (haut) et représentation annuelle des positions des images (bas) pour les années 1990, 1991, 1992, et 1993

Dans cette section, nous avons présenté une méthode permettant d'étudier la structure d'un média indépendamment de son contenu et nous l'avons testée sur les corpus de GDL et JDG. La section suivante présente des statistiques de base sur ces corpus en termes de données textuelles et n-grammes extraits sur la base des résultats de l'OCR.

5.3 Prétraitements des données

Les données textuelles ont été extraites de fichiers XML contenant les tokens reconnus pour chaque article ainsi que les métadonnées concernant les différentes parutions. Nous avons observé que, dans certain cas, le processus d'OCR retranscrit des caractères du journal en caractères spéciaux et non alphabétiques. C'est notamment le cas pour le caractère "l" parfois retranscrit comme un "|" ou bien le caractère "a" parfois retranscrit comme un "@". De plus, un caractère spécial peut aussi apparaître dans les données si une tache d'encre est présente sur la page scannée. L'évaluation de la qualité de l'OCR nous a permis de quantifier certains types d'erreurs et leurs évolutions dans le temps. Ainsi nous avons pu détecter des périodes bruitées. Une façon typique d'éliminer une grande partie du bruit est d'appliquer un filtre de fréquence avant toute analyse des données. Toutefois, nous verrons au cours de cette thèse que le filtre fréquentiel n'est pas toujours la meilleure méthode pour éliminer le bruit et nous préférons définir de nouvelles méthodes plus résistantes au bruit et aux erreurs d'OCR.

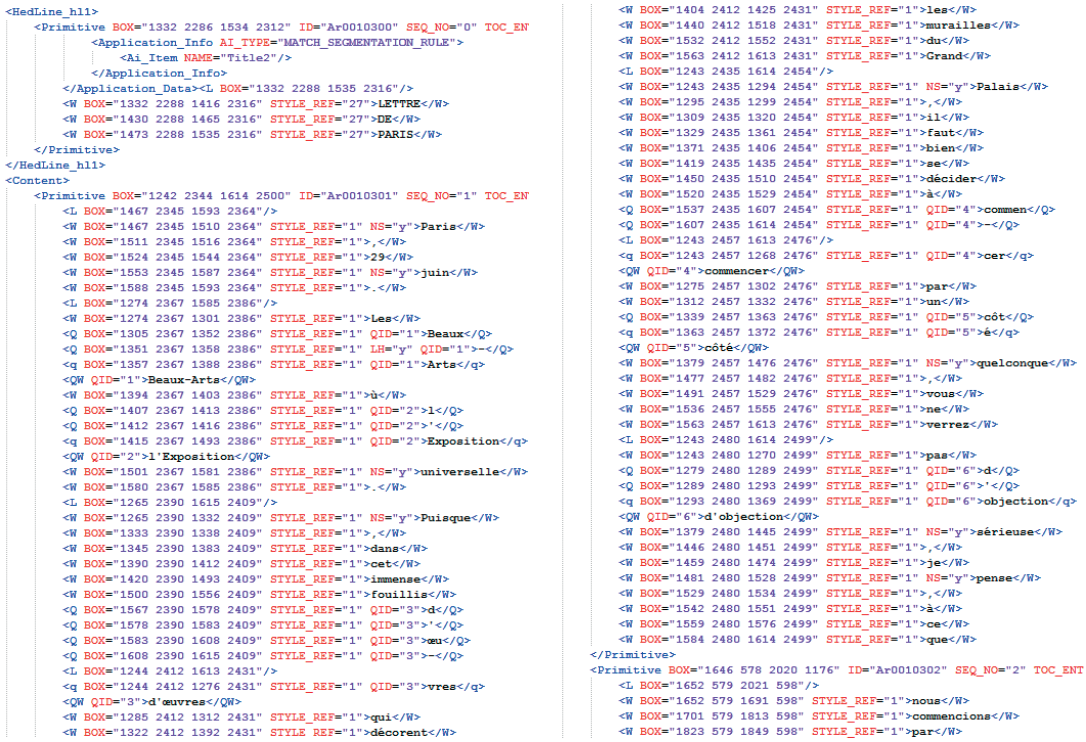
Dans toute étude de corpus, il est d'usage d'appliquer un prétraitement des données afin d'enlever les erreurs manifestes et obtenir un corpus de meilleurs qualité. Toutefois, afin de limiter la diversité des choix de prétraitement trop souvent arbitraires, nous optons pour trois différents prétraitements dont la méthodologie est simple et suffisamment efficace.

Les prétraitements sont les suivants :
- Données brutes : les mots sont extraits sans aucun changement et restent exactement les mêmes que ceux reconnus par le processus OCR initial.
- Données "alpha" : seuls les caractères alphabétiques sont reconnus, ignorant les caractères numériques et les caractères spéciaux.
- Données alphanumériques : seuls les caractères alphabétiques ou numériques sont acceptés, ignorant les caractères spéciaux.

Aucune différence n'est faite sur la base de la casse et les caractères majuscules sont transformées en minuscules. De plus, nous considérons que deux mots séparés par un tiret sont extraits des fichiers XML comme deux mots consécutifs différents. En effet, les tirets signifiant la séparation d'un unique mot sur deux lignes sont théoriquement reconnus et le mot est donc reconstitué dans les données initiales. Cela signifie toutefois que selon ce choix, certaines entités à mots composés se retrouveront décomposées comme les prénoms "Jean-François" qui sera traité comme l'entité "Jean" suivi de l'entité "François". De la même façon nous considérons les apostrophes également comme une séparation entre deux mots. Pour cette thèse et sans indication contraire, nous utilisons implicitement le prétraitement alphanumérique afin de pouvoir également analyser le comportement et l'évolution des chiffres et nombres dans le corpus. Pour étudier l'évolution diachronique du langage sur ces corpus, nous comparons les données de ceux-ci sur la base d'une période annuelle. Cela nous permet d'obtenir un équilibre entre le manque d'un nombre suffisant de points de comparaison et le fait d'avoir trop de fluctuations dans nos données, parce qu'elle sont trop peu significatives. Par exemple, le choix d'une période de 10 ans ne fournit au maximum que 20 points à comparer tandis qu'au

contraire, le choix d'une période mensuelle fournit suffisamment de points de comparaison, mais ces données ont tendance à fluctuer avec le temps, car elles sont peu représentatives. Une période annuelle fourni une balance entre ces deux effets.

La Figure 5.14 montre un exemple de fichier XML contenant les tokens et les métadonnées.



```
<HedLine_hl1>
    <Primitive BOX="1332 2286 1534 2312" ID="Ar0010300" SEQ_NO="0" TOC_EN
        <Application_Info AI_TYPE="MATCH_SEGMENTATION_RULE">
            <Ai_Item NAME="Title2"/>
        </Application_Info>
    </Application_Data><L BOX="1332 2288 1535 2316"/>
    <W BOX="1332 2288 1416 2316" STYLE_REF="27">LETTRE</W>
    <W BOX="1430 2288 1465 2316" STYLE_REF="27">DE</W>
    <W BOX="1473 2288 1535 2316" STYLE_REF="27">PARIS</W>
</Primitive>
</HedLine_hl1>
<Content>
    <Primitive BOX="1242 2344 1614 2500" ID="Ar0010301" SEQ_NO="1" TOC_EN
    <L BOX="1467 2345 1593 2364"/>
    <W BOX="1467 2345 1510 2364" STYLE_REF="1" NS="y">Paris</W>
    <W BOX="1511 2345 1516 2364" STYLE_REF="1">,</W>
    <W BOX="1524 2345 1544 2364" STYLE_REF="1">29</W>
    <W BOX="1553 2345 1587 2364" STYLE_REF="1" NS="y">juin</W>
    <W BOX="1588 2345 1593 2364" STYLE_REF="1">.</W>
    <L BOX="1274 2367 1585 2386"/>
    <W BOX="1274 2367 1301 2386" STYLE_REF="1">Les</W>
    <Q BOX="1305 2367 1352 2386" STYLE_REF="1" QID="1">Beaux</Q>
    <Q BOX="1351 2367 1358 2386" STYLE_REF="1" LH="y" QID="1">-</Q>
    <q BOX="1357 2367 1388 2386" STYLE_REF="1" QID="1">Arts</q>
    <QW QID="1">Beaux-Arts</QW>
    <W BOX="1394 2367 1403 2386" STYLE_REF="1">ù</W>
    <Q BOX="1407 2367 1413 2386" STYLE_REF="1" QID="2">l</Q>
    <Q BOX="1412 2367 1416 2386" STYLE_REF="1" QID="2">'</Q>
    <q BOX="1415 2367 1493 2386" STYLE_REF="1" QID="2">Exposition</q>
    <QW QID="2">l'Exposition</QW>
    <W BOX="1501 2367 1581 2386" STYLE_REF="1" NS="y">universelle</W>
    <W BOX="1580 2367 1585 2386" STYLE_REF="1">.</W>
    <L BOX="1265 2390 1615 2409"/>
    <W BOX="1265 2390 1332 2409" STYLE_REF="1" NS="y">Puisque</W>
    <W BOX="1333 2390 1338 2409" STYLE_REF="1">,</W>
    <W BOX="1345 2390 1383 2409" STYLE_REF="1">dans</W>
    <W BOX="1390 2390 1412 2409" STYLE_REF="1">cet</W>
    <W BOX="1420 2390 1493 2409" STYLE_REF="1">immense</W>
    <W BOX="1500 2390 1556 2409" STYLE_REF="1">fouillis</W>
    <Q BOX="1567 2390 1578 2409" STYLE_REF="1" QID="3">d</Q>
    <Q BOX="1578 2390 1583 2409" STYLE_REF="1" QID="3">'</Q>
    <Q BOX="1583 2390 1608 2409" STYLE_REF="1" QID="3">œu</Q>
    <Q BOX="1608 2390 1615 2409" STYLE_REF="1" QID="3">-</Q>
    <L BOX="1244 2412 1613 2431"/>
    <q BOX="1244 2412 1276 2431" STYLE_REF="1" QID="3">vres</q>
    <QW QID="3">d'œuvres</QW>
    <W BOX="1285 2412 1312 2431" STYLE_REF="1">qui</W>
    <W BOX="1322 2412 1392 2431" STYLE_REF="1">décorent</W>
    <W BOX="1404 2412 1425 2431" STYLE_REF="1">les</W>
    <W BOX="1440 2412 1518 2431" STYLE_REF="1">murailles</W>
    <W BOX="1532 2412 1552 2431" STYLE_REF="1">du</W>
    <W BOX="1563 2412 1613 2431" STYLE_REF="1">Grand</W>
    <L BOX="1243 2435 1614 2454"/>
    <W BOX="1243 2435 1294 2454" STYLE_REF="1" NS="y">Palais</W>
    <W BOX="1295 2435 1299 2454" STYLE_REF="1">,</W>
    <W BOX="1309 2435 1320 2454" STYLE_REF="1">il</W>
    <W BOX="1329 2435 1361 2454" STYLE_REF="1">faut</W>
    <W BOX="1371 2435 1406 2454" STYLE_REF="1">bien</W>
    <W BOX="1419 2435 1435 2454" STYLE_REF="1">se</W>
    <W BOX="1450 2435 1510 2454" STYLE_REF="1">décider</W>
    <W BOX="1520 2435 1529 2454" STYLE_REF="1">à</W>
    <Q BOX="1537 2435 1607 2454" STYLE_REF="1" QID="4">commen</Q>
    <Q BOX="1607 2435 1614 2454" STYLE_REF="1" QID="4">-</Q>
    <L BOX="1243 2457 1613 2476"/>
    <q BOX="1243 2457 1268 2476" STYLE_REF="1" QID="4">cer</q>
    <QW QID="4">commencer</QW>
    <W BOX="1275 2457 1302 2476" STYLE_REF="1">par</W>
    <W BOX="1312 2457 1332 2476" STYLE_REF="1">un</W>
    <Q BOX="1339 2457 1363 2476" STYLE_REF="1" QID="5">côt</Q>
    <q BOX="1363 2457 1372 2476" STYLE_REF="1" QID="5">é</q>
    <QW QID="5">côté</QW>
    <W BOX="1379 2457 1476 2476" STYLE_REF="1" NS="y">quelconque</W>
    <W BOX="1477 2457 1482 2476" STYLE_REF="1">,</W>
    <W BOX="1491 2457 1529 2476" STYLE_REF="1">vous</W>
    <W BOX="1536 2457 1555 2476" STYLE_REF="1">ne</W>
    <W BOX="1563 2457 1613 2476" STYLE_REF="1">verrez</W>
    <L BOX="1243 2480 1614 2499"/>
    <W BOX="1243 2480 1270 2499" STYLE_REF="1">pas</W>
    <Q BOX="1279 2480 1289 2499" STYLE_REF="1" QID="6">d</Q>
    <Q BOX="1289 2480 1293 2499" STYLE_REF="1" QID="6">'</Q>
    <q BOX="1293 2480 1369 2499" STYLE_REF="1" QID="6">objection</q>
    <QW QID="6">d'objection</QW>
    <W BOX="1379 2480 1445 2499" STYLE_REF="1" NS="y">sérieuse</W>
    <W BOX="1446 2480 1451 2499" STYLE_REF="1">,</W>
    <W BOX="1459 2480 1474 2499" STYLE_REF="1">je</W>
    <W BOX="1481 2480 1528 2499" STYLE_REF="1" NS="y">pense</W>
    <W BOX="1529 2480 1534 2499" STYLE_REF="1">,</W>
    <W BOX="1542 2480 1551 2499" STYLE_REF="1">à</W>
    <W BOX="1559 2480 1576 2499" STYLE_REF="1">ce</W>
    <W BOX="1584 2480 1614 2499" STYLE_REF="1">que</W>
</Primitive>
<Primitive BOX="1646 578 2020 1176" ID="Ar0010302" SEQ_NO="2" TOC_ENT
    <L BOX="1652 579 2021 598"/>
    <W BOX="1652 579 1691 598" STYLE_REF="1">nous</W>
    <W BOX="1701 579 1813 598" STYLE_REF="1">commencions</W>
    <W BOX="1823 579 1849 598" STYLE_REF="1">par</W>
```


FIGURE 5.14 – Exemple de fichier XML d'un article de GDL du 30 juin 1900

Nous observons dans la Figure 5.14 que la primitive "Ar0010302" a été reconnue comme la continuation de "Ar0010301" et constitue donc un seul article. Certains mots a priori considérés comme des unités séparées sont regroupés ensemble a posteriori comme par exemple les unités "Beaux", "-" et "Arts" pour former le token "Beaux-Arts". Toutefois, en vertu des règles de prétraitement choisies, ce mot sera de nouveau séparé en deux mots distincts par notre algorithme de prétraitement.

Pour le cas du token "d'oeuvre" la reconstitution est faite à partir des éléments détectés "d", "'", "oe" et "uvre". Ensuite, l'algorithme de prétraitement donnera la séparation "d" et "oeuvre" qui est plus adéquate pour l'étude linguistique, car les deux entités gardent leur rôles distincts au sein de la langue, ce qui n'aurait pas été le cas avec le token "d'oeuvre".

Les tokens "commen", "-" et "cer" ont également été rassemblés en le token "commencer" qui est un mot correspondant au sens de la phrase et qui est inclus dans le lexique du français (contrairement à "commen" et "cer"). La raison de la séparation initiale de ces tokens par un tiret est la césure d'un mot trop long en fin de ligne.

6 Statistiques

6.1 Statistiques de base

Cette section présente l'évolution annuelle de diverses statistiques de base comme le nombre de parutions, de pages par parution, de mots ou de mots par article. Le nombre de parutions par année est présenté dans la Figure 6.1.

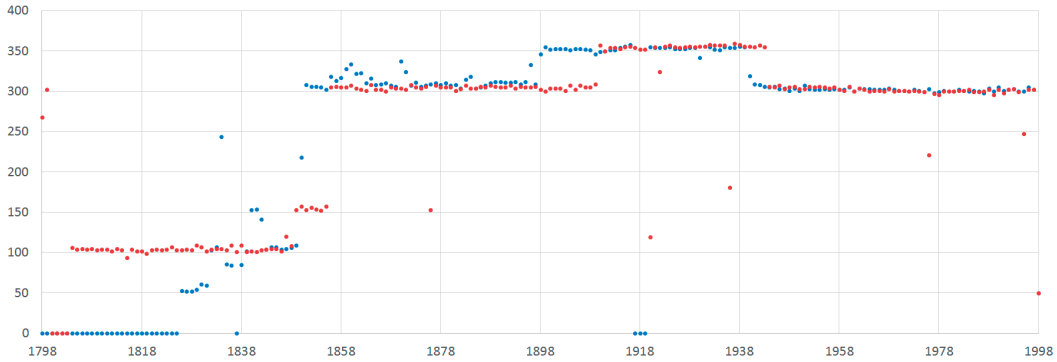


FIGURE 6.1 – Nombre de parutions par année pour JDG (bleu) and GDL (rouge)

Nous observons que le nombre de parutions varie par étapes successives. Le corpus de GDL commence avec deux années de parutions, mais s'arrête en 1800 avant de recommencer en 1804 sous le nouveau nom de "Gazette de Lausanne". Le nombre de parutions par année est plus faible dans les années les plus anciennes. Cependant, c'est le corpus de JDG qui atteint une publication annuelle régulière d'environ 300 par année dès 1851 suivi par le corpus de GDL en 1856. Ensuite, les deux journaux ont un nombre similaire de parutions par année avec des variabilités dues aux diverses parutions les weekends ou à des numéros spéciaux (principalement entre 1896 et 1939 pour JDG et entre 1910 et 1943 pour GDL). Pour les deux corpus, certaines années n'ont pas de données ou sont incomplètes. Nous présentons le nombre de pages par parution dans la Figure 6.2.

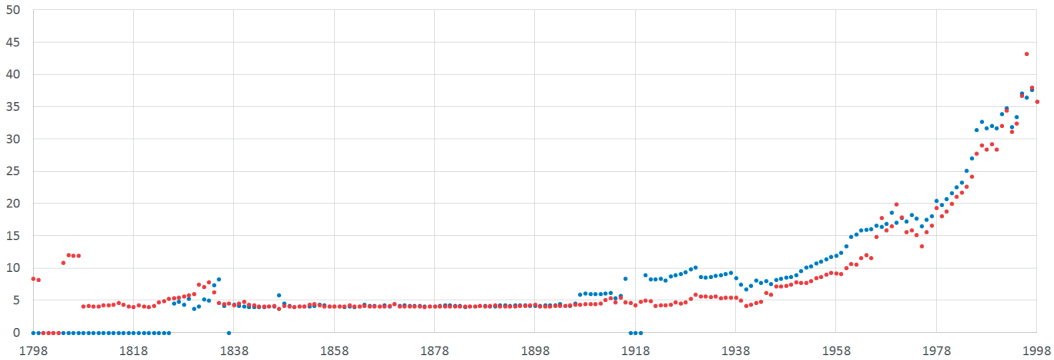


FIGURE 6.2 – Nombre de pages par parution par année pour JDG (bleu) et GDL (rouge)

Nous observons dans la Figure 6.2 que le nombre de pages par parution est stable autour de 4 pages jusqu'à 1907 pour GDL et 1945 pour JDG avec quelques exceptions avant 1838 pour les deux journaux. Après 1945, le nombre de pages augmente significativement, atteignant 30 à 40 pages au cours des dernières années. Ce nombre croissant de pages s'explique par le nombre croissant de suppléments et numéros spéciaux.

Nous comptons le nombre de n-grammes apparaissant chaque année avec les trois prétraitements définis précédemment. Ces valeurs ont ensuite été stockées dans une base de données MySQL afin de les indexer et les rendre consultables par requêtes de type SQL. Il est alors facile de calculer les fréquences des n-grammes en divisant leur nombre d'occurrences pour une année donnée par le nombre total de n-grammes annuels.

Dans l'extraction de données avec prétraitement alphanumérique, nous comptons approximativement 2.4 milliards de mots dans les corpus, 1.4 milliards pour JDG et 1 milliard pour GDL. Nous comptons également environ 14.5 millions de mots uniques pour JDG et 7.5 millions pour GDL. Il est intéressant de noter que la quantité globale de mots augmente quasiment de façon monotone croissante au cours des années, comme le montre la Figure 6.3.

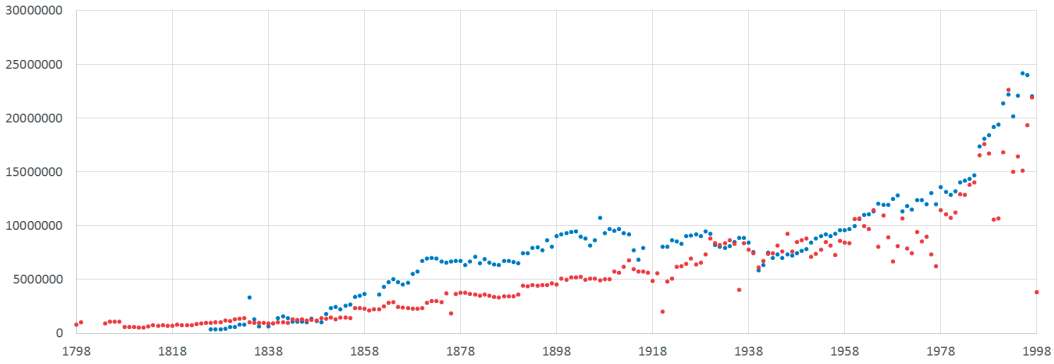


FIGURE 6.3 – Nombre de mots par année pour JDG (bleu) et GDL (rouge)

Cette propriété doit être considérée afin de ne pas constituer un biais dans la comparaison des différentes années. Il est toutefois intéressant de constater une différence importante du nombre de mots par année entre les deux journaux GDL et JDG. Cet effet est notable entre 1860 et 1930, incluant la période durant laquelle les deux journaux ont un nombre de quatre pages par parution. La taille du journal de JDG en terme de mots augmente progressivement, mais distancie celle de GDL qui par ailleurs augmente aussi. Nous présentons l'évolution de la taille moyenne des articles (nombre de mots par article pour chaque année) dans la Figure 6.4.

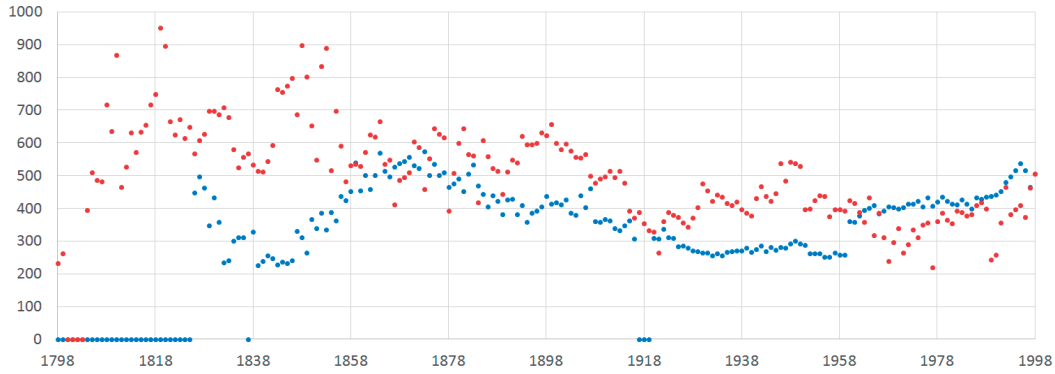


FIGURE 6.4 – Nombre de mots par article pour JDG (bleu) et GDL (rouge)

Nous observons dans la Figure 6.4 que le nombre de mots par article est assez instable pour GDL et tend à diminuer avec le temps. Cependant, JDG montre une image plus stable, bien que non dépourvue de variations et tendant à diminuer jusqu'en 1960, où la taille des articles augmente de nouveau au cours des dernières années. Sur la base des Figures 6.3 et 6.4, nous résumons certaines propriétés importantes de l'ensemble de données :

– Le corpus de GDL commence en 1798 avec peu de données. Les années 1800, 1801, 1802 et 1803 n'ont d'ailleurs aucune donnée. Cela nous amène dans certains cas à ne considérer les données de GDL qu'à partir de 1804.
– La dernière année, 1998, n'a pas la même quantité de données que les autres dernières années. Cela s'explique par la date de la fin des archives le 28 février 1998, de sorte que seuls deux mois sont représentés en 1998.
– Certaines valeurs aberrantes sont détectées notamment pour les années où des données sont partiellement manquantes comme en 1876, 1920 et 1936 pour GDL. Ces années devront donc être considérées avec prudence.
– Une valeur aberrante est observée pour l'année 1834 pour JDG. En effet, un bond impressionnant du nombre de mots est détecté. Cette année particulière est sujette à de nombreuses erreurs d'OCR et devra également être considérée avec prudence.
– Le nombre de mots par année est globalement plus instable pour GDL, particulièrement entre 1965 et 1977 et à partir de 1989.
– Le nombre de mots par articles varie de façon plus importante pour GDL, mais le corpus de JDG accuse un saut brusque et conséquent en 1960.

Ces statistiques de base nous permettent de détecter des années qui risquent de poser problème dans nos analyses. En outre, nous apprenons que la taille des données évolue de façon monotone croissante avec le temps. Cet effet pourrait causer un biais dans l'analyse diachronique de l'évolution linguistique des corpus. Ces constatations nous amènent à considérer les données avec prudence dans l'élaboration de nos méthodes d'analyse diachronique de corpus afin d'éviter des biais potentiels.

6.2 Statistiques fréquentielles

Dans cette section, nous regardons le comportement de la fréquences des mots sur l'entièreté des corpus de GDL et JDG. En 1935, George Kingsley Zipf observa une loi empirique (Zipf, 1935; Piantadosi, 2014) régissant les fréquences de mots dans divers corpus textuels. En classant les mots par fréquence descendante, il remarqua que le rang d'un mot multiplié par sa fréquence est une valeur à peu près constante. La loi de Zipf indique donc que la fréquence des mots dans un corpus suit une distribution particulière où la fréquence peut être exprimée par $f(r) = \frac{K}{r}$ avec r le rang du mot, $f(r)$ étant sa fréquence et K une constante.

Il est donc intéressant de représenter les n-grammes au travers de l'espace de leurs fréquences en fonction du rang afin de vérifier si ceux-ci suivent bien la loi de Zipf. Pour observer les différences entre les niveaux n des n-grammes, nous représentons la distribution fréquentielle des n-grammes en fonction du rang pour les corpus de JDG et GDL dans les Figures 6.6 et 6.5.

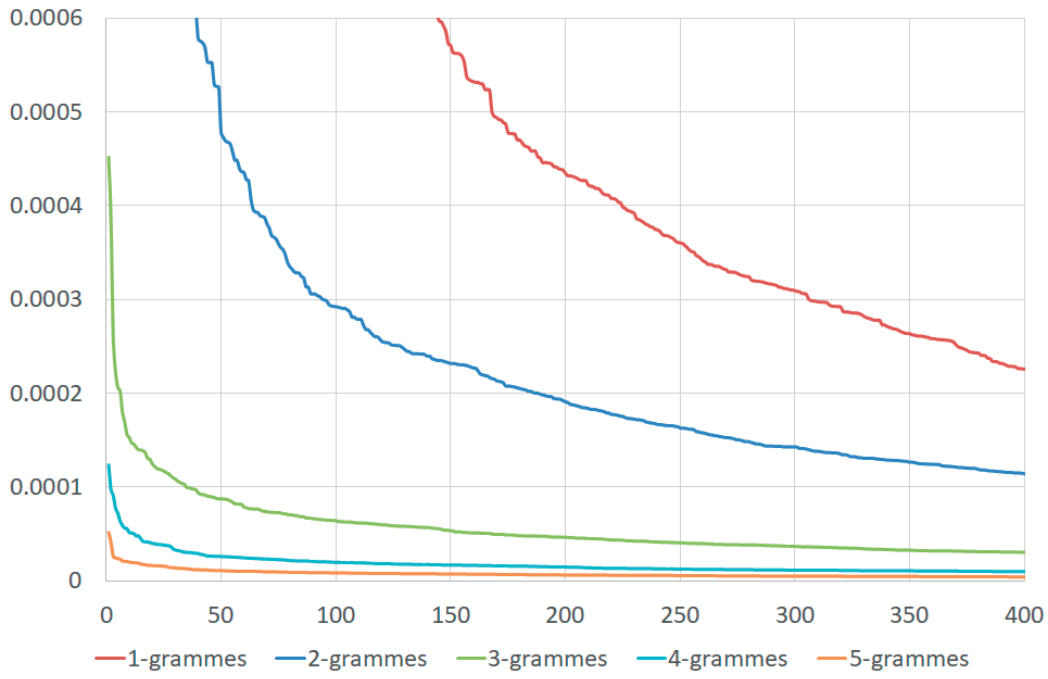


FIGURE 6.5 – Fréquence des n-grammes en fonction de leur rang pour GDL

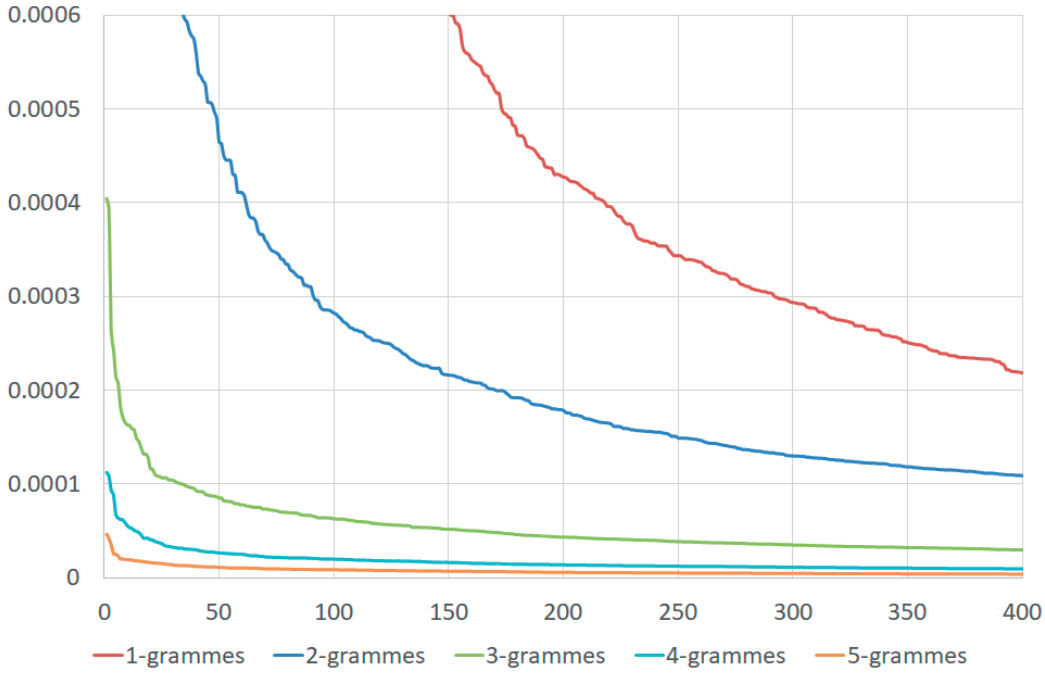


FIGURE 6.6 – Fréquence des n-grammes en fonction de leur rang pour JDG

Afin de vérifier si les n-grammes suivent la loi de Zipf d'une manière plus fiable, nous pouvons représenter les fréquences des n-grammes par rapport à l'inverse du rang de ces derniers. Tous les points devraient former une droite dont le coefficient angulaire est donné par la constante de Zipf. Ce graphe est présenté pour les mots dans la Figure 6.7.

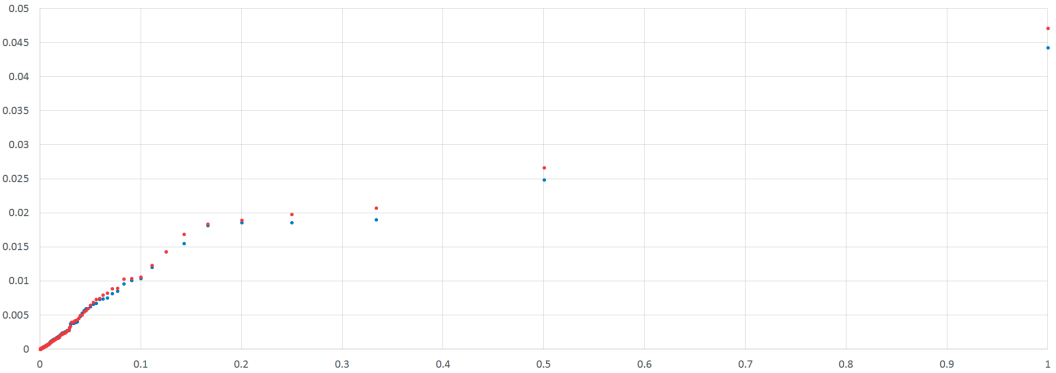


FIGURE 6.7 – Fréquence des mots en fonction de l'inverse de leur rang pour les corpus de JDG (bleu) et GDL (rouge)

Afin de mieux visualiser l'information que contient chaque point représenté, nous présentons les 100 mots les plus fréquents pour chaque corpus dans la table 6.1.

JDG						GDL					
rang	**id**	**freq**	**rang**	**id**	**freq**	**rang**	**id**	**freq**	**rang**	**id**	**freq**
1	de	4.42E-02	51	mais	2.03E-03	1	de	4.71E-02	51	sa	1.98E-03
2	la	2.49E-02	52	8	2.00E-03	2	la	2.67E-02	52	ses	1.87E-03
3	l	1.90E-02	53	elle	1.92E-03	3	le	2.07E-02	53	8	1.74E-03
4	à	1.86E-02	54	t	1.89E-03	4	à	1.98E-02	54	je	1.73E-03
5	le	1.86E-02	55	10	1.84E-03	5	l	1.89E-02	55	lui	1.67E-03
6	et	1.81E-02	56	y	1.83E-03	6	et	1.84E-02	56	i	1.65E-03
7	les	1.55E-02	57	50	1.82E-03	7	les	1.69E-02	57	ou	1.63E-03
8	d	1.43E-02	58	7	1.76E-03	8	d	1.43E-02	58	ces	1.62E-03
9	des	1.20E-02	59	6	1.71E-03	9	des	1.23E-02	59	être	1.61E-03
10	a	1.04E-02	60	sa	1.70E-03	10	du	1.06E-02	60	tout	1.60E-03
11	du	1.01E-02	61	ou	1.69E-03	11	a	1.04E-02	61	leur	1.60E-03
12	en	9.61E-03	62	ses	1.67E-03	12	en	1.03E-02	62	deux	1.57E-03
13	que	8.55E-03	63	j	1.58E-03	13	que	8.98E-03	63	vous	1.55E-03
14	un	8.18E-03	64	30	1.58E-03	14	un	8.86E-03	64	si	1.55E-03
15	il	7.53E-03	65	genève	1.54E-03	15	il	8.27E-03	65	comme	1.54E-03
16	une	7.40E-03	66	leur	1.52E-03	16	une	7.98E-03	66	fait	1.51E-03
17	est	7.34E-03	67	si	1.52E-03	17	est	7.43E-03	67	même	1.45E-03
18	qui	6.76E-03	68	ces	1.51E-03	18	qui	7.33E-03	68	était	1.43E-03
19	s	6.58E-03	69	suisse	1.50E-03	19	dans	6.89E-03	69	suisse	1.43E-03
20	dans	6.26E-03	70	être	1.48E-03	20	pour	6.43E-03	70	ils	1.41E-03
21	pour	6.00E-03	71	comme	1.46E-03	21	au	5.94E-03	71	6	1.40E-03
22	m	5.99E-03	72	20	1.45E-03	22	s	5.71E-03	72	avait	1.39E-03
23	au	5.67E-03	73	tout	1.44E-03	23	par	5.50E-03	73	h	1.31E-03
24	par	5.33E-03	74	fait	1.41E-03	24	qu	5.07E-03	74	e	1.31E-03
25	qu	4.82E-03	75	deux	1.41E-03	25	m	4.97E-03	75	0	1.30E-03
26	n	4.54E-03	76	lui	1.41E-03	26	on	4.55E-03	76	conseil	1.29E-03
27	on	4.07E-03	77	25	1.40E-03	27	n	4.35E-03	77	7	1.28E-03
28	l	3.98E-03	78	9	1.35E-03	28	ce	4.25E-03	78	bien	1.27E-03
29	ce	3.89E-03	79	ils	1.35E-03	29	se	4.16E-03	79	j	1.26E-03
30	pas	3.85E-03	80	conseil	1.34E-03	30	ne	4.02E-03	80	tous	1.24E-03
31	ne	3.84E-03	81	je	1.31E-03	31	sur	3.99E-03	81	sans	1.23E-03
32	se	3.83E-03	82	même	1.30E-03	32	pas	3.96E-03	82	après	1.22E-03
33	sur	3.77E-03	83	0	1.28E-03	33	plus	3.39E-03	83	où	1.21E-03
34	plus	3.19E-03	84	bien	1.28E-03	34	l	3.10E-03	84	10	1.17E-03
35	5	2.77E-03	85	fr	1.25E-03	35	ont	2.82E-03	85	dont	1.16E-03
36	nous	2.75E-03	86	sans	1.22E-03	36	avec	2.77E-03	86	t	1.15E-03
37	c	2.73E-03	87	avait	1.21E-03	37	son	2.67E-03	87	lausanne	1.11E-03
38	i	2.70E-03	88	00	1.20E-03	38	été	2.63E-03	88	30	1.10E-03
39	ont	2.57E-03	89	15	1.20E-03	39	nous	2.52E-03	89	faire	1.10E-03
40	avec	2.55E-03	90	12	1.19E-03	40	aux	2.43E-03	90	50	1.08E-03
41	3	2.55E-03	91	était	1.19E-03	41	cette	2.37E-03	91	20	1.08E-03
42	2	2.45E-03	92	tous	1.16E-03	42	5	2.35E-03	92	9	1.05E-03
43	4	2.43E-03	93	dont	1.11E-03	43	sont	2.35E-03	93	encore	1.03E-03
44	été	2.43E-03	94	r	1.09E-03	44	elle	2.34E-03	94	00	1.01E-03
45	e	2.39E-03	95	p	1.09E-03	45	c	2.31E-03	95	grand	1.01E-03
46	son	2.38E-03	96	vous	1.09E-03	46	3	2.21E-03	96	contre	9.87E-04
47	aux	2.31E-03	97	après	1.08E-03	47	mais	2.19E-03	97	fr	9.86E-04
48	h	2.27E-03	98	où	1.04E-03	48	2	2.19E-03	98	dit	9.78E-04
49	sont	2.22E-03	99	11	1.04E-03	49	4	2.15E-03	99	leurs	9.58E-04
50	cette	2.13E-03	100	faire	1.03E-03	50	y	2.04E-03	100	sous	9.48E-04

TABLE 6.1 – Les 100 mots les plus fréquents avec leur fréquence et leur rang pour les corpus de JDG (gauche) and GDL (droite)

Nous observons dans la Figure 6.7 que les cinq premiers points (fréquence supérieure) ne sont pas alignés. Cependant, en supprimant ces cinq premiers points (correspondant aux mots "de", "la", "l", "à" et "le"), un bon alignement visuel est constaté pour les 9995 autres points. Le même graphe excluant les 5 mots les plus fréquents est présenté dans la Figure 6.8.

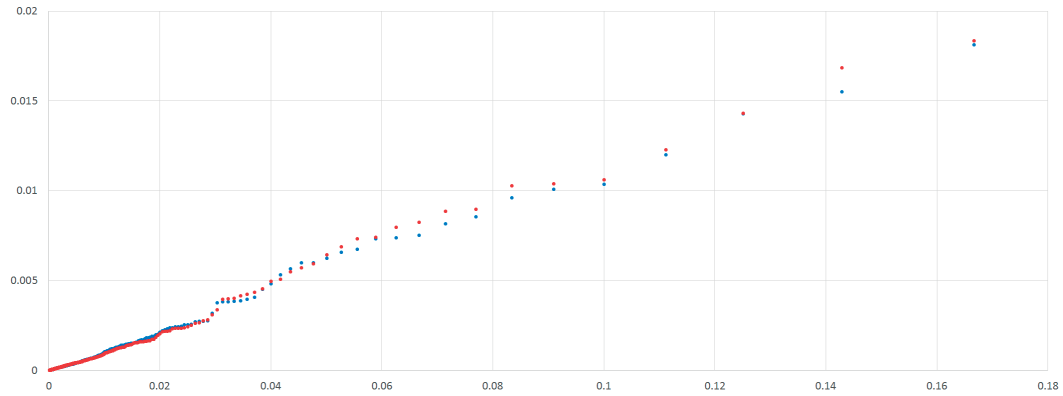


FIGURE 6.8 – Fréquence des mots par rapport à l'inverse des rangs des mots des corpus de JDG (bleu) et GDL (rouge) en excluant les 5 mots les plus fréquents

En étudiant les 10 000 mots les plus fréquents du corpus de JDG et GDL, nous pouvons ajuster la loi de Zipf sur la courbe observée en utilisant l'optimisation classique des moindres carrés. Compte tenu de la formule de Zipf, le processus d'optimisation est facilement obtenu par un calcul exact. Soit f_i la fréquence observée avec i allant de 1 au nombre de mots total, 10 000 dans notre exemple). Le résidu total R peut être exprimé comme :

$$R = \sum_i \left(f_i - f(r_i)\right)^2 = \sum_i \left(f_i - \frac{K}{r_i}\right)^2$$

En minimisant cette quantité, nous résolvons l'équation

$$\frac{\partial R}{\partial K} = 0 \;\Leftrightarrow\; -2\sum_i \left(f_i - \frac{K}{r_i}\right)\frac{1}{r_i} = 0 \;\Leftrightarrow\; \sum_i \left(\frac{f_i}{r_i}\right) = K\sum_i \left(\frac{1}{r_i^2}\right) \;\Leftrightarrow\; K = \frac{\sum_i \left(\frac{f_i}{r_i}\right)}{\sum_i \left(\frac{1}{r_i^2}\right)}$$

Une fois tous les paramètres trouvés, nous choisissons un critère de similarité afin de mesurer la qualité de l'ajustement. Nous comparons donc la courbe originale et celle de la loi de Zipf, ajustée en fonction du coefficient K, à l'aide de la similarité cosinus (Salton et McGill, 1986; Singhal, 2001). Cette mesure est par définition incluse dans l'intervalle [0,1] puisque toutes les fréquences sont positives. La similarité entre la loi de Zipf et la loi empirique observée est de 0.87 pour le corpus de JDG et 0.88 pour le corpus de GDL. Il faut noter que la similarité cosinus est indépendante de toute valeur multiplicative appliquée sur l'ensemble des fréquences et que donc la mesure ne dépend pas du processus d'optimisation qui détermine la constante de Zipf. Elle peut donc être calculée sans la détermination de la constante K.

Cependant, la loi de Zipf est une loi empirique et sa formulation n'est pas mathématiquement correcte, ainsi doit-elle être utilisée avec précaution. Une généralisation de la loi de Zipf a été faite en 1965 par le mathématicien Mandelbrot (Mandelbrot, 1965) qui présente une loi statistique formelle. Cependant, en 1964, le linguiste Gustav Herdan a proposé une amélioration de la loi de Zipf en formulant la loi de Herdan (Herdan, 1964). Cette loi partage son nom avec la loi de Heaps (Heaps, 1978), formulée d'un point de vue informationnel, car il a été démontré que les deux lois sont des formulations différentes du même phénomène (Egghe, 2007).

Nous testons ensuite la loi Zipf avec la distribution fréquentielle des n-grammes avec n allant de 1 à 5. Il ressort que cette loi est relativement bien adaptée pour $n = 1$ et $n = 2$, mais pas pour les valeur de n allant de 3 à 5. Fait intéressant, la loi de Heaps $f(r) = \frac{K}{r^t}$ semble bien correspondre à la distribution fréquentielle des n-grammes pour n allant de 1 à 5. La similarité doit être calculée après le processus d'optimisation, car la similarité cosinus n'est pas indépendante du paramètre t. Les résultats sont affichés dans la Table 6.2.

	JDG		**GDL**	
	Zipf(K)	**Heaps(K,t)**	**Zipf(K)**	**Heaps(K,t)**
1-grammes	0.8673	0.9503	0.8789	0.9531
2-grammes	0.8371	0.9906	0.8339	0.9914
3-grammes	0.3475	0.9782	0.3512	0.9798
4-grammes	0.3103	0.9807	0.3074	0.9832
5-grammes	0.2745	0.9795	0.2742	0.9784

TABLE 6.2 – Similarité cosinus pour les lois de Zipf et Heaps

Nous observons dans la Table 6.2 que si la qualité d'ajustement de la loi de Zipf diminue avec n, la qualité d'ajustement de la loi de Heaps garde des valeurs stables et supérieures à 95% pour tous n allant de 1 à 5. La recherche de (Petersen *et al.*, 2012) indique également que la loi statique de Zipf ne correspond qu'à un régime particulier concernant les mots fréquents et qu'un second régime existe concernant les mots nouveaux et moins fréquents. Cette dynamique s'explique par le fait qu'une langue qui se développe en terme lexical de façon importante subit un effet "refroidissant" ralentissant le "besoin" de nouveaux mots.

Cette section nous a permis de vérifier si nos données suivent la célèbre loi empirique de Zipf. Le contraire aurait évidement été étonnant, mais nous avons poussé l'investigation jusqu'aux n-grammes de niveaux $n < 6$. Depuis Zipf, plusieurs améliorations de la loi ont été proposées et nous avons constaté que la loi de Heaps généralise mieux le comportement des n-grammes que la loi de Zipf pour tous les niveaux $n < 6$. La prochaine section présente des observations statistiques empiriques diverses effectuées sur les corpus de GDL et JDG afin d'observer de façon descriptive le comportement des données textuelles de ces corpus.

6.3 Statistiques exploratoires diverses

Nous introduisons l'aspect diachronique des corpus en subdivisant ceux-ci en sous-corpus annuels et en considérant la loi statistique suivie par les fréquences de n-grammes pour chaque année. Nous déterminons la constante du Zipf pour chacune de ces années pour les 1-grammes et 2-grammes par la méthode d'optimisation des moindres carrés. Nous pouvons étudier l'évolution de la constante estimée de Zipf représentée dans les Figures 6.9 et 6.10 afin d'extraire de premières informations sur le comportement diachronique du corpus.

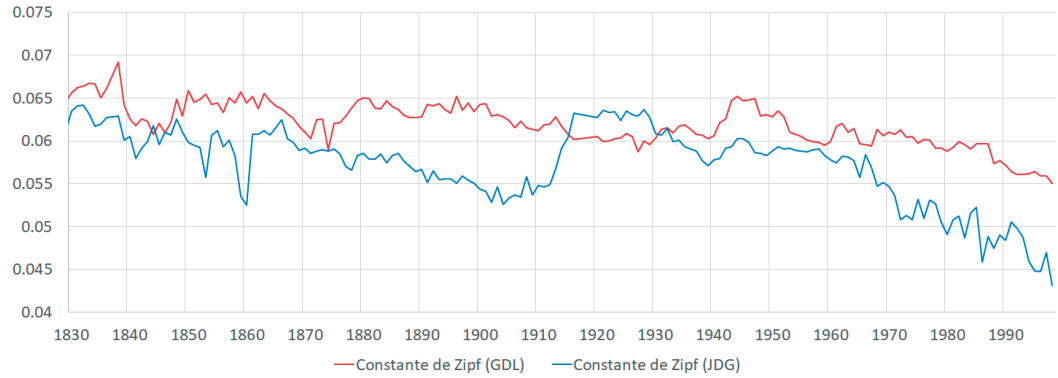


FIGURE 6.9 – L'évolution de la constante de Zipf avec les années pour les 1-grammes pour GDL (rouge) et JDG (bleu) avec le prétraitement alphanumérique

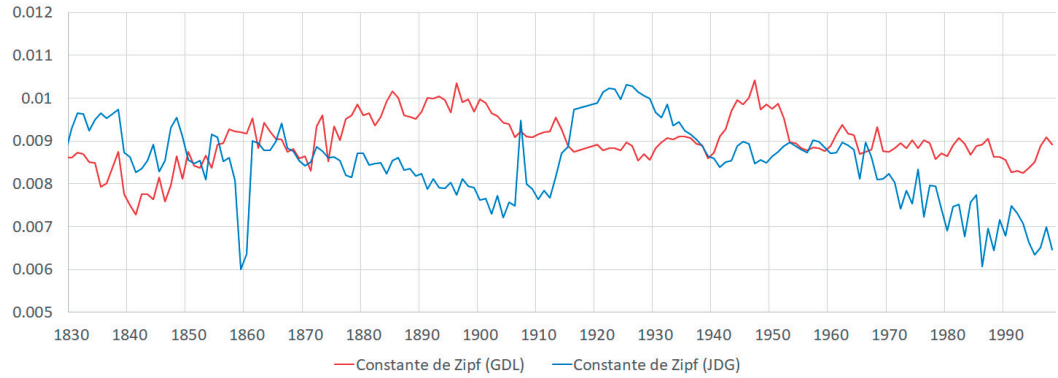


FIGURE 6.10 – L'évolution de la constante de Zipf avec les années pour les 2-grammes pour GDL (rouge) et JDG (bleu) avec le prétraitement alphanumérique

Nous constatons que la constante de GDL est plus stable que celle de JDG. Nous observons également que la constante de GDL est supérieure à celle de JDG, sauf entre 1915 et 1930. La constante de JDG semble diminuer particulièrement pour la période après 1970. Cet effet est visible pour les deux journaux, mais est plus marqué pour JDG.

Il est difficile d'interpréter ces valeurs sans autres indicateurs comparables, mais nous notons que le comportement global de ces courbes pourraient être corrélé avec l'évolution de la taille des corpus (en termes de nombre total de mots). Nous calculons une corrélation de Pearson entre la constante de Zipf des 1-grammes et la taille des corpus de -0.74 pour JDG et -0.84 pour GDL. Pour l'analyse des 2-grammes nous observons un comportement similaire, mais avec des valeurs extrêmes comme 1859, 1860 et 1907 pour JDG. Nous notons également que les années avant 1870 montrent un comportement différent entre l'analyse de 1-grammes et 2-grammes. Nous calculons une corrélation de Pearson entre la constante de Zipf des 2-grammes et la taille des corpus de -0.53 pour JDG et -0.52 pour GDL qui est inférieure aux corrélations calculées pour l'analyse des 1-grammes. Cependant, afin de vérifier que l'évolution des données numériques n'interfère pas avec la détermination des constantes de Zipf, nous calculons la même évolution statistique en utilisant le prétraitement alpha plutôt que l'alphanumérique. L'évolution de la constante Zipf estimée avec le prétraitement alpha est représentée dans les Figures 6.11 et 6.12.

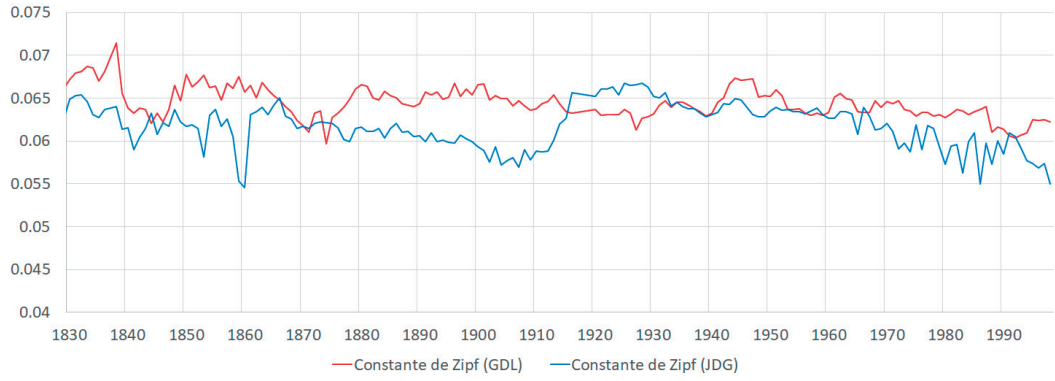


FIGURE 6.11 – L'évolution de la constante de Zipf avec les années pour les 1-grammes pour GDL (rouge) et JDG (bleu) avec le prétraitement alpha

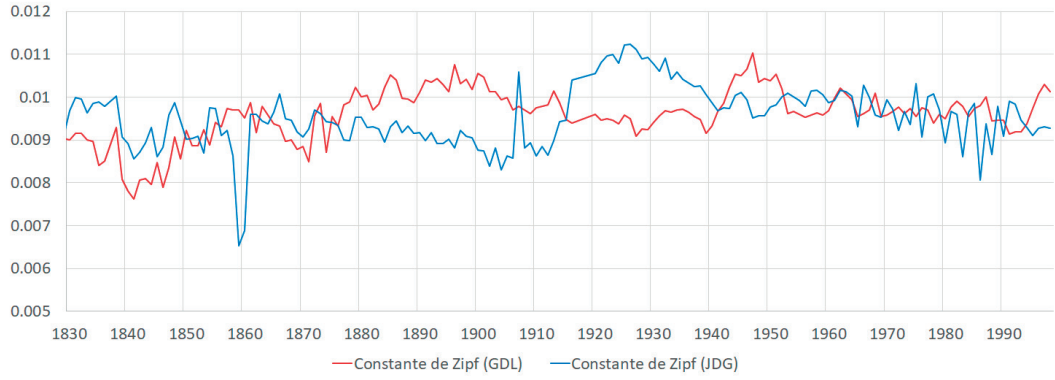


FIGURE 6.12 – L'évolution de la constante de Zipf avec les années pour les 2-grammes pour GDL (rouge) et JDG (bleu) avec le prétraitement alpha

Nous observons globalement le même comportement pour les deux types de prétraitement si ce n'est pour les années postérieures à 1970 où les constantes de Zipf sont significativement différentes. Les données numériques influencent donc l'estimation de la constante de Zipf. La corrélation globale entre la constante de Zipf et la taille du corpus est égale à -0.30 pour (JDG / 1-grammes), 0.16 (JDG / 2-grammes), -0.66 (GDL / 1-grammes) et 0.20 (GDL / 2-grammes). Il est toutefois possible que les variations des constantes de Zipf ne soient pas seulement causées par les variations de taille de corpus, mais aussi par des effets d'évolution linguistique.

Il est pourtant impossible d'émettre une hypothèse d'évolution linguistique sur la base de ces seuls graphes, car nous observons que la variation de la taille du corpus affecte des mesures simples comme l'évolution de la constante de Zipf. L'étude de l'évolution de celle-ci selon deux prétraitements différents nous amène également à considérer l'évolution des données numériques dans le texte.

Il nous faut donc développer des méthodes robustes dont la sensibilité à l'évolution de la taille du corpus est réduite tout comme la sensibilité à l'évolution des données numériques. Cela nous amène naturellement à étudier le comportement des données numériques seules. Nous nous intéressons à toutes les chaînes de caractères qui sont composées uniquement par des chiffres et qui décrivent donc des nombres au sein du corpus.

En 1938, Frank Benford observe une loi empirique qu'il appelle la loi des nombres anormaux (Benford, 1938) sur la fréquence du premier chiffre composant les nombres d'un corpus donné. Il observe non seulement que plus le chiffre est petit et plus il fréquent, mais également que cette fréquence peut se décrire par une équation mathématique simple comme la suivante :

$$f(d) = log_{10}\left(1 + \frac{1}{d}\right)$$

où d est la décimale allant de 1 à 9. Cela signifie que la fréquence observée empiriquement dans la plupart des corpus du monde "réel" vaut respectivement environ 30.1%, 17.6%, 12.5%, 9.7%, 7.9%, 6.7%, 5.8%, 5.1% et 4.6% pour les chiffres 1 à 9. Aujourd'hui, cette loi empirique est utilisée dans la détection de divers type de fraudes comme la fraude fiscale, comptable, électorale, scientifique, etc. Toutefois, il existe des valeurs "corrigées" qui s'appliquent selon le type de corpus utilisé. Ainsi, nous utiliserons également les valeurs empiriques observées dans le cadre de corpus de presse afin de les comparer à celle obtenues pour GDL et JDG.

La comparaison de la loi empirique de Benford, de la loi de Benford corrigée pour les corpus de presse et de nos observations sur JDG et GDL est présentée dans la Figure 6.13. Nous y observons que la distribution de JDG et GDL est similaire à la loi de type "journal" de Benford à l'exception des fréquences des chiffres "4" et "5" qui ont des rôles totalement inversés. Il est possible que cette inversion soit due au fait de l'apparition dans les années les plus récentes de sections boursières ainsi que d'horaire de bus et de trains, contenant de nombreuse données numériques et favorisant les occurrences du chiffre "5" au détriment du "4".

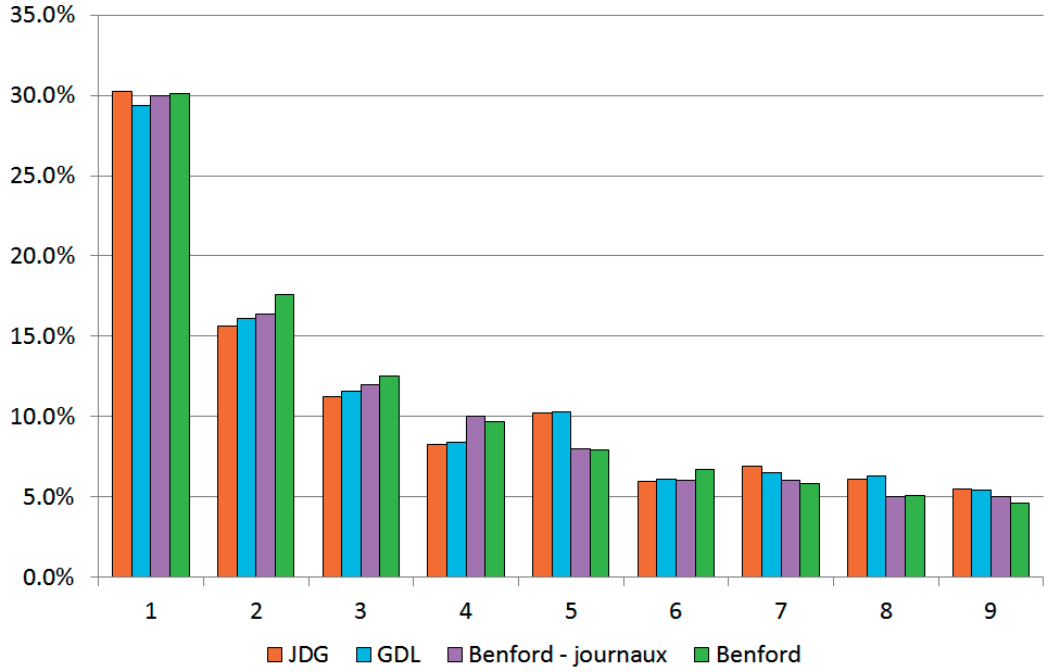


FIGURE 6.13 – La loi de Benford (vert), loi de type "journal" de Benford (violet), la loi d'apparition de premier chiffre des corpus de JDG (orange) et GDL (bleu)

En considérant la fréquence d'apparition des nombres à part entière, nous observons aisément que les données numériques suivent également une loi de Zipf. Nous observons qu'en règle générale, un nombre est d'autant plus fréquent qu'il représente une valeur faible. Nous tentons donc une analyse de la distribution fréquentielle des nombres en fonction non pas de leur rang, mais d'eux-mêmes. Ainsi, il est intéressant de savoir si la valeur de la fréquence d'un nombre multiplié par ce nombre lui-même est constant dans un corpus donné. Le graphe de la fréquence des nombres en fonction d'eux-mêmes est présentée dans la Figure 6.14.

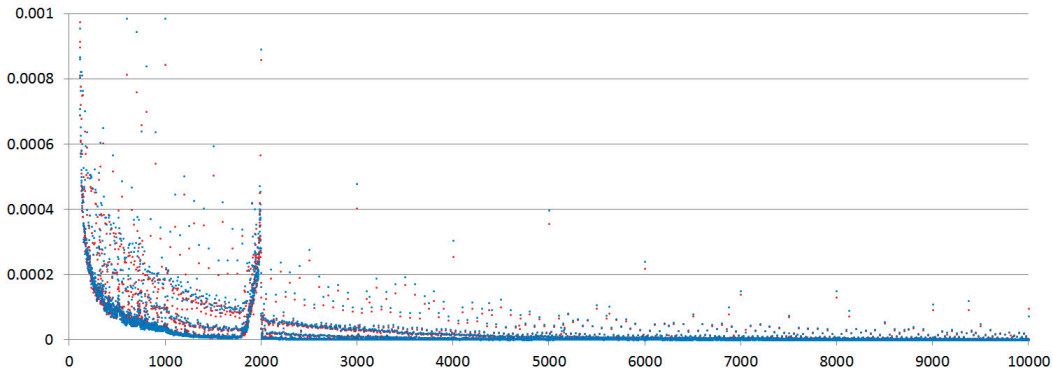


FIGURE 6.14 – Distribution fréquentielle des nombres dans GDL (rouge) et JDG (bleu)

Ce graphe est similaire à ceux obtenu sur d'autres corpus (Delahaye et Gauvrit, 2013) et nous observons de la même façon la pointe se dessinant entre les années 1800 et 2000 en raison de la signification particulière supplémentaire de ces nombres qui sont également des années. Nous observons plusieurs régimes selon les derniers chiffres composant le nombre. Nous identifions les catégories suivantes :

– Catégorie 1 : nombres finissant par 000.
– Catégorie 2 : nombres finissant par 500.
– Catégorie 3 : nombres finissant par 00 et ne faisant pas partie de la catégorie 1.
– Catégorie 4 : nombres finissant par 25.
– Catégorie 5 : nombres finissant par 50.
– Catégorie 6 : nombres finissant par 75.
– Catégorie 7 : nombres finissant par 0 et ne faisant pas partie des catégories 1 ou 3.
– Catégorie 8 : nombres finissant par 5 et ne faisant pas partie des catégories 4 ou 6.
– Catégorie 9 : nombres ne faisant pas partie des catégories 1 à 8.

Chacune de ces catégories semble obéir à sa propre lois de Zipf ou de Heaps. Nous ajustons donc les lois de Zipf et Heaps sur chacune de ces catégories et mesurons la qualité de l'ajustement grâce à la similarité cosinus. Les résultats sont présentés dans la Table 6.3. Nous observons dans la Table 6.3 que la loi Heaps n'améliore pas de façon importante la similarité cosinus. Cependant, la décomposition des données numériques en ces 9 catégories améliore sensiblement la similarité cosinus pour chacune des catégories, ce qui confirme l'existence de plusieurs régimes de fréquences numériques selon cette loi de Zipf modifiée (loi Zipf dont on remplace le rang par le nombre lui-même).

	JDG		**GDL**	
	Zipf(C)	**Heaps(C,t)**	**Zipf(C)**	**Heaps(C,t)**
Toutes catégories	0.6540	0.7695	0.6619	0.7768
Catégorie 1	0.9048	0.9293	0.9295	0.9490
Catégorie 2	0.9873	0.9994	0.9870	0.9992
Catégorie 3	0.9537	0.9660	0.9628	0.9678
Catégorie 4	0.7794	0.9524	0.7876	0.9326
Catégorie 5	0.8784	0.9952	0.8999	0.9909
Catégorie 6	0.9727	0.9870	0.9733	0.9814
Catégorie 7	0.8420	0.8438	0.7980	0.8054
Catégorie 8	0.9859	0.9861	0.9897	0.9931
Catégorie 9	0.7947	0.8826	0.7852	0.8752

TABLE 6.3 – Similarité cosinus pour les loi de Zipf et Heaps sur les catégories numériques des corpus de GDL et JDG

6.4 Synthèse

Ce chapitre a présenté diverses observations statistiques sur les corpus qui constituent notre matière première dans ce travail de thèse. Certaines observations ont un impact important sur la suite de celle-ci tandis que d'autres ont simplement un objectif exploratoire.

Nous faisons un résumé de ces diverses observations :
- La taille du corpus en terme de nombre de mots augmente avec les années.
- La taille moyenne des articles en terme de nombre de mots varie au fil du temps et cela peut avoir une incidence sur la mesure des diversités lexicales et évolution linguistique.
- La distribution fréquentielle des mots suit la loi de Zipf.
- La distribution fréquentielle des n-grammes ne suit pas la loi de Zipf dès $n = 3$.
- La distribution fréquentielle des n-grammes de niveaux $n < 6$ suit la loi de Heaps.
- Les premiers chiffres dans chaque nombre suivent la loi de Benford (améliorée pour les corpus de presse) sauf les fréquences des chiffres "4" et "5" qui s'inversent approximativement.
- Les nombres suivent la loi de Zipf.
- Les nombres suivent une loi de Zipf modifiée (remplaçant le rang de ceux-ci par eux-mêmes), mais suivant des régimes différents en fonction des derniers chiffres composant le nombre.
- La distribution fréquentielle des mots sur un sous-corpus annuel suit la loi de Zipf.
- La constante de Zipf pour chaque sous-corpus annuels n'est pas constante dans le temps, mais son évolution est corrélée avec celle de la taille des corpus.

Les premiers résultats d'une analyse diachronique simple au travers de l'évolution de la constante de Zipf sont impactés par l'évolution de la taille du corpus. Toutefois, ils montrent des similitudes et des différences dans les corpus de GDL et JDG et doivent être comparés à d'autres indicateurs à développer dans le but d'éliminer les biais et cibler plus précisément l'évolution linguistique excluant le bruit autant que possible.

Dans cette thèse, nous utilisons la fréquence des n-grammes comme une mesure de base pour analyser les changements linguistiques. Aussi la prudence est de mise quant au prétraitement des données ainsi que les évolutions parallèles du corpus (nombres, sujets, nouvelles sections, erreurs d'OCR, etc) pouvant donner naissance à des artefacts et de fausses interprétations.

La partie suivante expose les concepts théoriques et méthodes que nous avons développés afin d'explorer les différents niveaux n des n-grammes des corpus de GDL et JDG tout en réduisant les effets du bruit et de la variation de taille de ces corpus.

Partie III Concepts et méthodes

7 Niveau Micro

Dans ce chapitre, nous présentons les concepts et méthodes de base liés à l'analyse diachronique des corpus à une échelle que nous appelons niveau Micro. A cette échelle, nous détectons les variations et les évolutions linguistiques au niveau des éléments de base du langage, des mots et des expressions, utilisés au cours du temps.

7.1 Profil fréquentiel

La notion de profil fréquentiel est un élément fondamental de l'étude de l'évolution linguistique sur des corpus de données textuelles. Cette notion est à la base de la plupart des analyses effectuées dans cette thèse. Le profil fréquentiel permet de décrire un aspect de l'évolution de l'utilisation d'un mot ou d'une expression en fonction du temps.

Définition 2. *Le profil fréquentiel d'un n-gramme dans un corpus donné est la série temporelle des fréquences relatives de ce n-gramme au sein du corpus selon une subdivision en plusieurs périodes de durées égales.*

Dans cette thèse le terme profil fréquentiel désigne par défaut un profil dont la période de subdivision est annuelle. A priori, rien n'empêche de choisir une période mensuelle ou par décades, mais le choix de périodes annuelles semble ici un bon compromis entre représentativité et nombre de fréquences décrivant le profil fréquentiel d'un n-gramme.

Cette définition nous indique la méthode de calcul des profils fréquentiels. Le corpus étudié est subdivisé en sous-corpus, typiquement selon les années de parution des journaux. Ensuite chacun des n-grammes est compté afin d'obtenir ce que l'on peut appeler une fréquence absolue d'occurrences du n-gramme considéré pour une année donnée. Ces fréquences sont ensuite normalisées, c'est-à-dire divisées par le nombre total de n-grammes comptés sur la même année. Nous obtenons une fréquence que l'on appelle fréquence relative, dont la valeur se situe entre 0 et 1 puisqu'un n-gramme particulier ne peut pas apparaître plus de fois que la totalité des n-grammes. La suite ordonnée dans le temps de ces fréquences relatives pour un n-gramme donné constitue son profil fréquentiel.

En analysant le profil fréquentiel de chaque n-gramme, nous pouvons explorer des changements linguistiques particuliers au niveau des entités de base que sont les mots et les combinaisons de mots. Le profil fréquentiel rend compte d'événements linguistiques particuliers comme par exemple l'apparition des mots et des expressions ainsi que leur disparition, tout comme des pics de fréquence importants au cours de certaines périodes, comme la tendance globale stable ou instable, à la hausse ou à la baisse de l'utilisation de certain n-grammes dans les corpus étudiés. Un exemple de profil fréquentiel, celui du 2-gramme "Conseil fédéral" dans le corpus JDG est présenté dans la Figure 7.1.

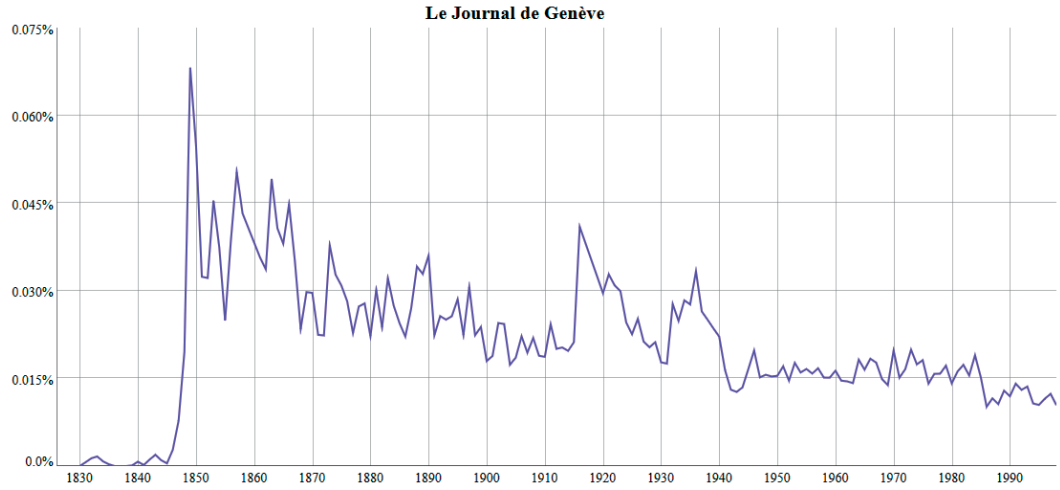


FIGURE 7.1 – Profil fréquentiel du 2-gramme "Conseil fédéral"

Le 2-gramme "Conseil fédéral" apparaît entre 1847 en 1849. L'apparition de ce 2-gramme dans les journaux suisses n'est pas anodin puisque le Conseil fédéral, organe exécutif de la Confédération suisse, a été créé en 1848 par la Constitution fédérale suisse. Cet exemple est donc plus spécifique au corpus et à l'histoire de la Suisse qu'à la langue. Nous observons que la fréquence relative du 2-gramme diminue légèrement avec le temps, mais se stabilise ensuite. Compte tenu de l'ordre de grandeur de son profil fréquentiel, "Conseil fédéral" se trouve parmi les 2-grammes les plus fréquents dans les deux corpus de JDG et GDL dès 1849.

Ces profils fréquentiels et leurs variations peuvent être une source d'information quant à l'évolution diachronique d'une expression donnée au sein du corpus, mais ils permettent aussi des comparaisons entre plusieurs profils fréquentiels de n-grammes ayant un lien spécifique (sémantique par exemple) ou même la comparaison de ceux-ci dans différents corpus. Les causes provoquant des variations de profils fréquentiels sont multiples. Elles peuvent être liées à des événements historiques, des annonces répétées, des sujets particulièrement à la mode ou même du bruit. Toutefois, il est aussi possible de trouver des exemples de causes liées à des changements linguistiques comme le remplacement sémantique d'un mot par un autre, une réforme de l'orthographe, un changement de sens d'un mot ou n-gramme particulier, etc.

Les études basées sur la visualisation des profils fréquentiels des n-grammes ont été populaires ces récentes années, notamment au travers des études du domaine "Culturomics" (Michel *et al.*, 2011; Delahaye et Gauvrit, 2013). Dans l'article (Michel *et al.*, 2011), les auteurs explorent la langue anglaise entre les années 1800 et 2000 en se focalisant sur les phénomènes linguistiques et culturels décrits par les profils fréquentiels des n-grammes provenant du corpus English Google Books. Ils étudient le cas de n-grammes particuliers et formulent des hypothèses culturelles et linguistiques basées sur les profils fréquentiels de ces n-grammes. Il est important de comprendre que ces profils fréquentiels n'ont pas vocation à prouver formellement une hypothèse linguistique ou culturelle, mais ils permettent néanmoins de tester ces hypothèses sur une base quantitative et peuvent, le cas échéant, constituer un ensemble d'indices permettant de supporter ou infirmer ces hypothèses.

Ce faisant, l'étude (Michel *et al.*, 2011) a ouvert une voie de recherche importante consistant à étudier et analyser les tendances culturelles et linguistiques au travers de grands corpus de textes numérisés. Ils ont montré l'apport de ces méthodes dans des domaines aussi variés que la lexicographie, l'évolution de la grammaire, la mémoire collective et les nouvelles technologies. Cette thèse utilise ce type de moyens d'analyses orientés big data sur les corpus journalistiques de JDG et GDL. Certaines méthodes sont également testées sur les corpus de Google Books, car ceux-ci sont plus grands et permettent d'explorer différentes langues. Cependant, les corpus de GDL et JDG ont l'avantage d'être clairement défini et délimité.

Dans le cas d'un n-gramme fréquent, le profil fréquentiel a l'avantage d'être représentatif, car le mot apparaît dans le corpus de façon suffisamment régulière. Dans le cas d'un n-gramme rare ou même unique, le profil fréquentiel n'est pas représentatif, car la probabilité d'apparition du n-gramme est faible et le corpus ne constitue plus un échantillon suffisamment grand que pour l'estimer. Il est pourtant clair que plus le niveau n augmente et plus un n-gramme aura tendance à devenir unique quel que soit le corpus.

Outre l'analyse classique des profils fréquentiels des n-grammes et leurs comparaisons, il est essentiel de comprendre les phénomènes de base qui sous-tendent les variations des ces fréquences relatives. Pour cela, il est possible de regrouper les profils fréquentiels en fonction de leurs typologies observées. Un exemple de quatre types de courbes fréquentielles de 1-grammes du corpus de JDG est présenté dans la Figure 7.2.

Sur cet exemple, nous observons les types suivants :
– Pic de fréquence asymétrique.
– Augmentation de fréquence périodique.
– Augmentation monotone continue.
– Forme d'un signal carré.

Il est possible de définir un nombre arbitraire de types du moment que suffisamment de courbes puissent se rattacher à l'un d'entre eux. Il est préférable que ces types représentent des ensembles ne se chevauchant pas, surtout si ceux-ci sont utilisés dans un but de classification.

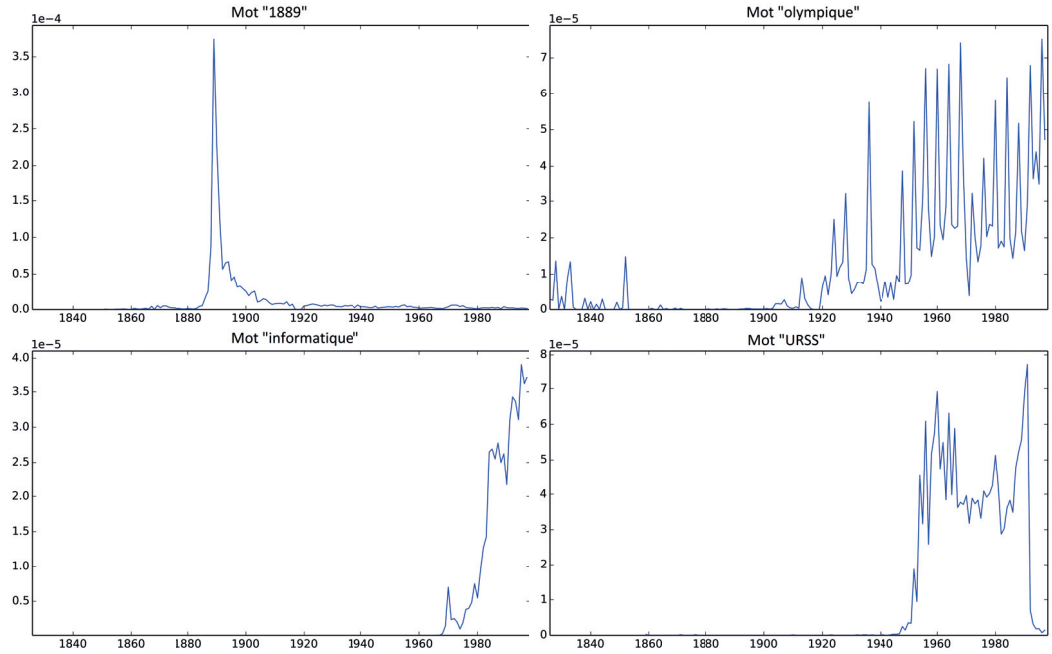


FIGURE 7.2 – Quatre exemples de courbes fréquentielles : "1889" (haut/gauche), "olympique" (haut/droite), "informatique" (bas/gauche) et "URSS" (bas/droite)

Au travers de ces différents types, nous observons que certains mots et expressions spécifiques semblent répondre à des lois générales ou à des modèles. Le mot "1889" (mot représentant le nombre 1889 et également la date 1889) est fréquent dans l'année 1889, mais cesse d'être intéressant assez rapidement. Toutefois une mémoire subsiste, le pic de fréquence est donc asymétrique. Le mot "olympique" augmente en fréquence de façon cyclique répétant le processus tous les 4 ans. Le mot "informatique" apparaît en 1968 et conserve une augmentation presque linéaire jusqu'à la limite du corpus en 1998. Le mot "URSS" n'est fréquent qu'entre 1950 et 1990, décrivant un signal carré avec une apparition abrupte, une stabilisation et pour finir un déclin tout aussi abrupt.

Nous avons observé que la majorité des mots représentant des dates couvertes par le corpus ont des profils fréquentiels au comportement similaire autant dans JDG que dans GDL, une apparition abrupte est observée à la date représentée par le mot en formant un pic important de fréquences et ensuite une diminution rapide, mais conservant une mémoire qui finit elle aussi par lentement disparaître avec le temps.

Dans cette thèse, le profil fréquentiel des n-grammes est un élément essentiel, car la plupart des méthodes, concepts et outils sont développés sur la base de ceux-ci. L'avantage de ce procédé est de permettre la réduction des hypothèses de modélisation à celles du profil fréquentiel. Par exemple, le fait que le profil fréquentiel est une notion indépendante de la langue entraine que les méthodes développées sur cette même base le sont également.

7.2 Corrélations diachroniques

Il est possible de comparer les profils fréquentiels deux à deux. Cette méthode est coûteuse, car non seulement le temps d'exécution, mais aussi le coût de stockage vont dépendre du nombre d'éléments de la matrice de comparaison, qui équivaut au carré du nombre de profils fréquentiels considérés. Toutefois, la méthode devient envisageable si nous restreignons l'espace des profils fréquentiels comparés aux plus intéressants. Cette méthode a l'avantage de ne faire aucune hypothèse sur la typologie existante au sein des courbes.

Une mesure intéressante permettant de comparer deux courbes est la corrélation. Nous utilisons pour cela le coefficient de corrélation Pearson, mais tout autre mesure de similarité est envisageable pour étudier la relation des courbes deux à deux.

Définition 3. *Le coefficient de corrélation linéaire R entre deux profils fréquentiels,* $\{G_i \mid i = 1, ..., n\}$ *de fréquence moyenne* $\bar{G}$ *et* $\{H_i \mid i = 1, ..., n\}$ *de fréquence moyenne* $\bar{H}$*, est égal au rapport de leur covariance et du produit de leurs écarts types, soit*

$$R = \frac{\sum_{i=1}^{n} \left(G_i - \bar{G}\right)\left(H_i - \bar{H}\right)}{\sqrt{\left(\sum_{i=1}^{n} \left(G_i - \bar{G}\right)^2\right)}\sqrt{\left(\sum_{i=1}^{n} \left(H_i - \bar{H}\right)^2\right)}}.$$

Cette mesure est comprise entre les valeurs -1 (maximalement corrélés négativement) et 1 (maximalement corrélés positivement). Plus cette mesure se rapproche de 0 et plus les profils fréquentiels seront faiblement corrélés.

Un avantage de cette mesure est son indépendance vis-à-vis de la multiplication d'une série fréquentielle par une constante. Celà signifie que la comparaison de deux courbes ayant une fréquence moyenne très différente ne sera pas "pénalisée", car la corrélation ne tient compte que de leurs variations vis-à-vis de leurs propres fréquences moyennes.

Cette propriété permet de découvrir des liens potentiels entre des n-grammes n'ayant pas forcément une fréquence moyenne proche, mais tout de même liés dans le sens qu'une augmentation proportionnelle de l'un impliquera une augmentation de même proportion de l'autre et vice-versa. Toutefois, il est établi que la corrélation entre deux séries temporelles n'implique pas immédiatement une causalité entre celles-ci, car une variable cachée pourrait également être la cause du même type d'effet.

Dans le cas particulier de la comparaison entre deux n-grammes contenant eux-mêmes un n-gramme commun de plus petite taille (par exemple "je joue du piano" et "je joue au football"), nous déduisons par la définition même du n-gramme et du profil fréquentiel, que la corrélation entre ces deux n-grammes sera en moyenne plus élevée que dans le cas de deux n-grammes ne contenant aucun élément en commun.

L'outil de comparaison des courbes deux à deux reste donc une méthode simple malgré son coût computationnel. Cette méthode s'affranchit de toute hypothèse sur les profils fréquentiels et peut être appliquée quel que soit les n-grammes considérés. Il en résulte une matrice carrée et symétrique dont le nombre de lignes et de colonnes est égal au nombre de n-grammes considérés. Cette matrice est en fait un espace de dissimilarités et peut être vue indirectement comme un espace de points dont les éléments les plus proches sont les plus corrélés. L'étude de cet espace permet potentiellement de révéler les liens qu'entretiennent les n-grammes en terme de comparaison de leurs profils fréquentiels.

7.3 Décomposition des profils fréquentiels

Certains profils fréquentiels de n-grammes ont des trajectoires indépendantes les uns des autres. Toutefois, d'autres sont au contraire dépendants et liés par une équation simple. il s'agit des n-grammes contenant un ou plusieurs n-grammes communs de plus petite taille. Cette relation traduit en fait un lien existant entre les différents niveaux n des n-grammes. Afin d'étudier cette relation, il est essentiel de revenir au modèle de base qui lie les profils fréquentiels des n-grammes avec ceux des (n+1)-grammes.

Afin de simplifier ce modèle, nous représentons les n-grammes comme une structure en arborescence selon une vision prospective. De chaque mot, partent diverses branches correspondant aux 2-grammes commençant par ce mot. Ensuite le même procédé est répété pour les niveaux $n > 1$ jusqu'au niveau maximum considéré, $n = 9$. Afin de caractériser formellement une manière de naviguer dans cette structure, nous définissons les notions suivantes :

Définition 4. *Le n-gramme g contient un (n+1)-gramme h si la séquence composée des n premiers mots de h est égale à g.*

Il peut paraître contre-intuitif d'exprimer qu'un n-gramme d'une certaine taille contient un (n+1)-gramme de taille plus grande. Toutefois, il faut raisonner en terme d'arborescence et de fréquence. Dans le cas de g qui contient h, il est clair que la fréquence de g ne peut pas être inférieure à celle de h. Par exemple, "je nage" ne peut pas avoir une fréquence inférieure à celle de "je nage rapidement" puisque "je nage" apparaît au moins en même temps que "je nage rapidement" alors que l'inverse n'est pas forcément vrai. Dans ce cas, la fréquence de h (contenu dans g) est vue comme l'une des composantes de la fréquence de g. Nous définissons également les termes "ascendant" et "descendant" :

Définition 5. *Le n-gramme g est un descendant du (n-1)-gramme h si g est contenu dans h.*

Définition 6. *Le n-gramme g est un ascendant du (n+1)-gramme h si g contient h.*

A titre d'exemple, le 2-gramme "la maison" est un descendant du mot "la" tandis que "la" est un ascendant de "la maison".

De ces définitions, découlent les propriétés suivantes :

Propriété 1. *Un n-gramme g de niveau $n > 1$ ne peut avoir qu'un seul ascendant, $A(g)$.*

Propriété 2. *Un n-gramme g a zéro, un ou plusieurs descendants, $D(g) = \{D(g)_t\}$ avec $t \in \mathbb{N}$ allant de 1 à $\|D(g)\|$.*

Il est possible d'exprimer le profil fréquentiel d'un n-gramme comme une décomposition en la somme de tous les profils fréquentiels de ses descendants. Soit un n-gramme donné g de profil fréquentiel $G = \{G^i \mid i = 1, ..., w\}$ et l'ensemble des descendants de g, $\{D(g)_t \mid t = 1, ..., v\}$. Le profil fréquentiel du (n+1)-gramme $D(g)_t$ est noté $D(G)_t = \{D(G)_t^i \mid i = 1, ..., w\}$.

Dans ces notations, la variable i allant de 1 à w représente le découpage en année du profil fréquentiel tandis que la variable t allant de 1 à v représente les descendants du n-gramme g. Si nous supposons qu'aucun article du corpus ne se termine par le n-gramme g, alors la relation suivante découle de la définition même du n-gramme et de son profil fréquentiel :

$$G_i = \sum_{t=1}^{v} D(G)_t^i \quad \forall i \in [1, ..., w]$$

Pour la suite, nous simplifions cette formulation de la façon suivante :

$$G = \sum_{t=1}^{v} D(G)_t$$

Si nous retirons l'hypothèse qu'aucun article du corpus ne se termine par le n-gramme g alors un terme de bord $R(G)$, indépendant des descendants de g et équivalent à la fréquence relative du n-gramme se trouvant en fin d'article, s'introduit dans l'équation. Nous avons donc la propriété suivante :

Propriété 3. *Soit un n-gramme g et son profil fréquentiel G, soit $D(G)_t$ les profils fréquentiels des descendants de g et soit $R(G)$ le terme de bord équivalent à la fréquence relative de g dans le cas où il se trouve en fin d'article, alors nous avons la propriété de décomposition suivante :*

$$G = \sum_{t=1}^{v} D(G)_t \; + \; R(G)$$

Le terme de bord $R(G)$ est relativement faible par rapport à la somme des profils fréquentiels des descendants de g, car les cas où g se trouve à la fin des articles est généralement plus rare que celui où g se trouve dans toutes les autres positions des articles. De plus, ce terme ne donne pas d'information quand à la relation entre les différents niveaux n. Nous négligeons donc ce terme dans la modélisation. Etant donné la nature récursive de la définition de "descendant", il est possible de décomposer le profil fréquentiel d'un n-gramme en la somme de tous les profils fréquentiels des (n+m)-grammes descendants par transitivité de ce n-gramme.

Nous définissons les notions de descendance et ascendance d'ordre m suivantes :

Définition 7. *Un descendant d'ordre m d'un n-gramme g, noté $D^m(g)$, donné est un (n+m)-gramme contenu dans le n-gramme g.*

Définition 8. *Un ascendant d'ordre m d'un n-gramme g, noté $A^m(g)$, donné pour $n > m$ est un (n-m)-gramme contenant le n-gramme g.*

En utilisant les propriétés de décomposition selon les descendants récursifs, nous pouvons écrire la propriété suivante :

Propriété 4. *Soit un n-gramme g et son profil fréquentiel G, soit $D^m(G)_t$ les profils fréquentiels des descendants d'ordre m de g et soit $R^m(G)$ le terme de bord d'ordre m, alors nous avons la propriété de décomposition d'ordre m suivante :*

$$G = \sum_{t_1} D(G)_{t_1} + R(G) = \sum_{t_2} D^2(G)_{t_2} + R^2(G) = \sum_{t_3} D^3(G)_{t_3} + R^3(G) = ... = \sum_{t_m} D^m(G)_{t_m} + R^m(G)$$

Le terme de bord d'ordre m, $R^m(G)$, est d'autant plus élevé que m est élevé, car il s'exprime comme la somme des termes de bord de chaque ordre précédent plus celui lié à l'ordre m lui-même. Toutefois, ce terme ne contient pas les informations propres à l'explication des profils fréquentiels, car sa détermination n'est liée qu'à sa position finale dans l'article. En le retirant du modèle, nous avons l'équation de décomposition suivante :

$$G \simeq \sum_{t_1} D(G)_{t_1} \simeq \sum_{t_2} D^2(G)_{t_2} \simeq \sum_{t_3} D^3(G)_{t_3} \simeq ... \simeq \sum_{t_m} D^m(G)_{t_m}$$

Cette propriété met en équation la relation existante entre les n-grammes se contenant. Elle permet de comprendre pourquoi ces n-grammes spécifiques ne sont pas indépendants.

En utilisant cette propriété, nous pouvons décomposer une courbe de profil fréquentiel d'un mot donné en la somme de tous les n-grammes descendants pour chaque niveau n donné. Toutefois, plus n est élevé, plus la valeur totale de la décomposition changera en raison des effets de bord plus importants. Cependant, nous pouvons tout de même expliquer une partie non négligeable du profil fréquentiel des mots en analysant tour à tour chaque niveau de décomposition des descendants.

Cette méthode permet donc de retracer la façon dont un mot a été utilisé comme point de départ afin de former des n-grammes spécifiques au cours du temps. Il est donc possible d'observer l'histoire de ce type d'utilisation au travers d'une décomposition différente pour chaque niveau n considéré. Une des faiblesses de la méthode est que la décomposition selon un niveau n va mettre en évidence tous les profils fréquentiels des n-grammes dont le premier élément est le mot initial considéré, y compris des profils potentiellement peu intéressants, peu représentatifs ou difficiles à interpréter.

7.4 Décomposition minimale des profils fréquentiels

La décomposition des profils fréquentiels permet d'étudier l'histoire d'un mot en le décomposant selon chaque niveau n désiré, mais elle ne permet pas de décomposer un profil fréquentiel en somme de n-grammes de longueur différente. Par exemple, nous considérons que le 7-gramme "il n'y a pas eu de" devrait être considéré de la même façon que le 4-gramme "il y a de". Pour cette raison, nous présentons une approche multi-échelle permettant de produire une analyse récursive pour différentes valeurs de n.

Intuitivement, l'objectif est de détecter le niveau d'autonomie de n-grammes particuliers, que l'on appelle des "expressions solidifiées", qui suivent des trajectoires évolutives simples et qui peuvent donc servir de base à la décomposition et à l'interprétation des évolutions complexes de l'utilisation de certain mots. Par exemple, prenons le mot "maison" dont la décomposition au niveau des 2-grammes inclus "maison blanche" et "maison de". Le 2-gramme "maison blanche" aura une trajectoire propre et relève d'une façon particulière d'utilisation le mot maison. Par contre, le 2-gramme "maison de" ne permet pas de spécifier suffisamment l'utilisation du mot "maison" et il serait alors intéressant de poursuivre la décomposition dans ce cas particulier.

Pour remédier à ce problème, nous regardons le processus de décomposition comme une arborescence dans laquelle il faut décider pour chacun des noeuds (n-grammes) si le profil fréquentiel doit être intégré dans la décomposition tel quel (selon un critère d'arrêt à déterminer) ou s'il faut continuer la décomposition au travers des descendants suivants afin de trouver des expressions plus intéressantes. L'idée est donc de parcourir l'arborescence branche par branche et décider s'il faut déployer un niveau supplémentaire de l'arbre à chacun des noeuds, ce qui revient à décider à partir de où on arrête la décomposition. Moyennant un critère d'arrêt raisonnable, cette méthode nous conduit à une décomposition multi-échelle des n-grammes. Divers critères d'arrêt peuvent être définis et nous optons pour un critère basé sur la modélisation du comportement des courbes fréquentielles.

Soit un modèle de courbe déterminé à l'avance, il est alors possible d'effectuer un fitting du modèle sur les profils fréquentiels "candidats" pour la décomposition. Nous utilisons alors une similarité de type cosinus (Salton et McGill, 1986; Singhal, 2001) afin de déterminer à quel point la courbe se comporte de façon similaire au modèle choisi. Dès que la similarité atteint un seuil fixé a priori, alors le profil fréquentiel du n-gramme est inclus dans la décomposition et il n'est donc pas nécessaire de continuer la décomposition au travers de ses descendants. Le n-gramme est alors appelé "expression solidifiée".

Nous appelons cette méthode, la décomposition minimale et celle-ci permet d'obtenir une décomposition multi-échelle d'un profil fréquentiel d'un mot ou d'une expression en tous ceux des n-grammes descendants qui se comportent de façon similaire au modèle choisi. L'étude de ces espaces de décomposition selon la décomposition minimale permet d'analyser l'évolution diachronique d'un mot ou d'une expression en retraçant son histoire sur la base de modèles et non pas d'une longueur arbitraire de n-grammes.

8 Niveau Macro

Dans ce chapitre, nous présentons les concepts de base que nous utilisons pour l'analyse diachronique des corpus au niveau Macro. Nous y détectons les variations et les évolutions globales en tentant de construire un indice numérique représentant l'ensemble des effets linguistiques en fonction du temps. La difficulté consiste à cibler l'évolution linguistique plus que celle d'autres propriétés du corpus comme par exemple la taille. Ces concepts ont également été présentés dans les articles (Buntinx *et al.*, 2017a) (Buntinx *et al.*, 2017b).

8.1 Distances et dissimilarités

La quantification des changements linguistiques dans les grands corpus est un problème largement abordé et ce domaine se développe de façon plus importante encore depuis la récente disponibilité de grandes bases de données textuelles. Une méthode couramment utilisée consiste à découper le corpus étudié en une série de sous-corpus représentant différentes périodes du corpus et établir une mesure de la distance textuelle ou dissimilarité entre chacun de ces sous-corpus afin d'étudier l'évolution de la mesure choisie.

Dans le travail de (Bochkarev *et al.*, 2014), les auteurs ont utilisé la divergence de Kullback-Leibler sous forme symétrique entre deux ensembles de fréquences de mots. Ils ont calculé cette mesure sur le corpus de Google Books (Michel *et al.*, 2011) afin de déterminer l'évolution lexicale dans plusieurs langues. D'autres études (Pechenick *et al.*, 2015b) (Pechenick *et al.*, 2015a) ont utilisé le corpus de Google Books et les divergences de Kullback-Leibler et Jensen-Shannon. Ils ont analysé les contributions spécifiques des mots les plus fréquents à la distance.

Un autre travail (Cocho *et al.*, 2015) a utilisé l'évolution de fréquence des mots et abordé l'analyse des changements linguistiques à travers le concept de diversité de la langue. Dans un travail récent, des physiciens et mathématiciens ont utilisé l'entropie généralisée sur des séquences de symboles avec une distribution de fréquence à queue longue (Gerlach *et al.*, 2016). Leur méthode est particulièrement adaptée à la répartition textuelle des mots dans un corpus suivant la loi empirique de Zipf (Zipf, 1935) (Piantadosi, 2014).

Une distance intuitive parmi les plus classiques est celle de Jaccard (Jaccard, 1901) (Jaccard, 1912). Cette distance a été découverte par Paul Jaccard afin d'estimer la diversité des espèces végétales entre deux échantillons. Dans notre cas, elle s'applique au lexique des sous-corpus de la subdivision périodique.

Soit deux corpus C_1 et C_2, et leurs lexiques (liste des mots uniques contenu dans le corpus), $L(C_1) \equiv L_1$ et $L(C_2) \equiv L_2$, la distance de Jaccard $d(L_1, L_2)$ est définie comme :

$$d(L_1, L_2) = 1 - \frac{|L_1 \cap L_2|}{|L_1 \cup L_2|} = 1 - \frac{|L_1 \cap L_2|}{|L_1| + |L_2| - |L_1 \cap L_2|}$$

de la même façon, d'autres indices peuvent également être utilisés comme par exemple la divergence de Kullback-Leibler (Kullback et Leibler, 1951) (Kullback, 1987), la distance Chi carré (Sakoda, 1981), ou la similarité cosinus (Salton et McGill, 1986; Singhal, 2001).

La distance de Jaccard est une mesure qui détermine la similarité de deux textes en fonction du nombre d'éléments lexicaux en commun. Cette distance est complémentaire à la notion de connexion lexicale (Muller, 1980) et est donc basée sur la présence ou l'absence des mots dans les lexiques, ignorant leurs fréquences d'occurrences.

La distance de Jaccard est sensible à l'apparition de nouveaux mots et la disparition d'anciens mots dans le corpus. Elle permet donc potentiellement de détecter les variations de vocabulaire et l'évolution de la diversité lexicale.

La distance de Jaccard est une métrique (Levandowsky et Winter, 1971) et satisfait donc les propriétés suivantes :

– Séparation : $d(L_1, L_2) = 0 \equiv L_1 = L_2$
– Symétrie : $d(L_1, L_2) = d(L_2, L_1)$
– Inégalité triangulaire : $d(L_1, L_3) \leqslant d(L_1, L_2) + d(L_2, L_3)$

Ces propriétés en font donc une distance au sens mathématique du terme. Toutefois, cette distance étant basée sur l'absence et la présence de mots, le bruit (erreurs d'OCR, évolution des sujets, pages de bourse, horaires de bus et de train) est un facteur qui peut affecter de façon importante la mesure de l'évolution linguistique.

Une façon courante de réduire l'effet du bruit est de fixer un seuil fréquentiel afin de filtrer l'ensembles des mots uniques en fonction de leur fréquence. Tous les mots dont la fréquence est en deçà du seuil choisi sont alors ignorés. Un inconvénient de cette méthode est le fait que la détermination du seuil est arbitraire et difficile à justifier.

Le calcul de cette distance sur l'ensemble des sous-corpus de la subdivision périodique détermine une matrice carré symétrique (ce qui n'est pas forcément le cas s'il s'agit d'une dissimilarité). Il est alors possible d'observer au travers de cette matrice à quel point la distance augmente quand les périodes considérées deviennent de plus en plus éloignées dans le temps.

8.2 Noyau et ensemble résilient

En plus du niveau de bruit, d'autres propriétés du corpus étudié peuvent perturber les résultats d'une analyse de dissimilarité classique et être interprétées comme une évolution linguistique. Par exemple, si la quantité annuelle de données analysées est dépendante du temps, comme nous le constatons d'ailleurs dans les corpus de JDG et GDL, alors la variation de taille du corpus affectera la plupart des mesures de dissimilarité utilisées.

Afin d'étudier globalement l'évolution des mots, nous avons choisi d'introduire la notion de noyau et plus généralement d'ensemble résilient permettant d'étudier l'évolution de l'ensemble des mots stables dans le corpus.

Définition 9. *Le noyau K_{t_1,t_2} est l'ensemble des mots uniques communs à toutes les subdivisions périodiques t du corpus C tel que $t_1 \leqslant t \leqslant t_2$.*

Soit L_t le lexique du sous-corpus correspondant à la période t, une définition équivalente est donnée par la formule suivante :

$$K_{t_1,t_2} = \bigcap_{t=t_1}^{t_2} L_t$$

Le noyau détermine donc un ensemble lexical stable sur une certaine durée déterminée par la différence entre la période t_2 et t_1. Nous définissons ensuite la notion de noyau résilient :

Définition 10. *Le noyau résilient K d'un corpus est le noyau $K_{t_{min},t_{max}}$ tel que t_{min} correspond à la première subdivision périodique du corpus tandis que t_{max} correspond à la dernière.*

Le noyau résilient correspond donc à la liste de mots uniques dont chaque élément apparaît une moins une fois dans chaque subdivision périodique du corpus. Il s'agit en fait des mots les plus stables du corpus. Etendant la notion de noyau résilient, nous définissons également l'ensemble résilient :

Définition 11. *L'ensemble résilient R_d est l'union de tous les noyaux $K_{x,y}$ correspondant à une durée de $y - x \geqslant d$.*

L'ensemble résilient contient donc tous les mots qui apparaissent un minimum de d périodes consécutives dans le corpus. A partir de la définition de l'ensemble résilient, nous définissons la résilience d'un mot :

Définition 12. *La résilience r d'un mot w dans le corpus C est donnée par la formule suivante : $r(w,C) = max\{d \mid w \subset R_{d,C}\}$.*

La résilience d'un mot est donc le nombre maximum de périodes consécutives durant lesquelles le mot apparaît dans le corpus. Cette propriété permet de sélectionner une partie des mots constituant le lexique du corpus non pas en fonction de la fréquence, mais plutôt de la stabilité et la persistance dans le corpus.

8.3 Distance nucléaire

Au lieu d'analyser les éléments qui changent rapidement dans le corpus, nous optons donc pour étudier les éléments les plus stables du lexique à travers les notions de résilience et de noyau résilient. La réduction de l'ensemble des mots analysés aux plus résilients nous permet d'exclure efficacement le bruit.

Le noyau résilient étant défini comme l'ensemble des mots communs à tous les sous-corpus de la subdivision périodique, il est clair que l'application d'une distance de type Jaccard à cet ensemble ne peut que donner qu'une valeur nulle, puisqu'on considère les même mots pour chaque subdivision périodique et que la distance de Jaccard ne considère que la présence ou l'absence d'éléments dans les ensembles comparés.

Nous introduisons donc une nouvelle distance permettant d'étudier l'ensemble des mots appartenant uniquement au noyau résilient en considérant un nouveau paramètre pour chaque sous-corpus : l'ordre du lexique en terme de fréquences. Nous définissons ensuite la distance nucléaire sur cette base :

Définition 13. *Soit le noyau résilient K et $I_t(w)$ l'indice du mot w dans K ordonné par les fréquences correspondant au sous-corpus de la période t, la distance nucléaire entre les périodes 1 et 2 est donnée par la formule suivante :*

$$d_{12}^{K} = \frac{1}{M} \sum_{w \subset K} |I_1(w) - I_2(w)|$$

Cette distance se normalise en la divisant par une valeur calculée comme la distance appliquée à deux ensembles de même taille que le kernel K, mais tel que l'ordre du second ensemble est l'inverse du premier. Soit $N = \|K\|$, la norme M vaut donc :

$$M = \sum_{i=1}^{N} |i - (N + 1 - i)| = \sum_{i=1}^{N} |2i - N + 1|$$

Par définition, la sensibilité de la mesure à la taille de l'ensemble des mots du corpus étudié est réduite, car l'ensemble des mots est réduit aux mots communs à tous les sous-corpus de la subdivision périodique et la fréquence relative d'un mot ne participe à la mesure qu'à la concurrence de son rang en comparaison des autres mots. Toutefois, une taille de sous-corpus trop réduite permet à la fréquence des mots, même appartenant au noyau, de varier avec plus d'ampleur et peut donc avoir un effet non négligeable sur la distance nucléaire.

En outre, cette distance ne s'encombre pas des mots apparaissant de manière plus ponctuelle dans le corpus et résiste donc à diverses évolutions non linguistiques comme les événements journalistiques et historiques ponctuels ou bien simplement à la diversité linguistique induite par l'évolution des sujets traités par le journal. Les principaux avantages de cette méthode sont donc de permettre la réduction significative du bruit, de cibler avec plus de pertinence l'évolution linguistique et de mieux résister aux effets de variations de la taille du corpus.

8.4 Entropie et entropie nucléaire

L'entropie est une mesure de la thermodynamique introduite par Rudolf Clausius (Clausius, 1868). La mesure d'entropie représente le degré de désorganisation du contenu informationnel d'un système. Claude Shannon (Shannon, 1948) a ensuite formalisé cette notion en terme de théorie de l'information au travers de la mesure de l'entropie de Shannon.

Cette mesure s'exprime de la façon suivante :

$$H = -\sum_{i=1}^{N} p_i \; ln(p_i)$$

avec p_i la probabilité d'apparition du n-gramme i, N le nombre de n-grammes dans le système. Nous utilisons ici le logarithme népérien et non pas le logarithme binaire. La propriété suivante est respectée :

$$\sum_{i=1}^{N} p_i = 1$$

Dans le cas des n-grammes, les probabilités sont estimées par le calcul de la fréquence relative de ceux-ci au sein du corpus.

La notion d'entropie exprime le degré d'information contenu dans un système et, de façon équivalente, sa diversité. A titre d'exemple, le travail (Juola, 2013) mesure la complexité de la culture par l'entropie de Shannon sur les 1-grammes et 2-grammes dans le corpus de Google Books. Celui-ci conclut que la culture se complexifie avec le temps. Toutefois, dans le cadre d'une étude de corpus diachronique, cette mesure est potentiellement corrélée avec l'évolution de la taille du corpus. Cette hypothèse se base sur le fait qu'un corpus de taille plus grande va permettre une diversité d'expression supérieure dans ce corpus. Cela entrainerait alors une augmentation de l'entropie due à l'augmentation de la taille corpus plutôt qu'à l'évolution de son contenu.

Afin de réduire cet effet, nous proposons de combiner la mesure de l'entropie de Shannon avec la notion de noyau résilient. Il suffit de renormaliser les fréquences sur les mots ou n-grammes contenus dans le noyau résilient et appliquer la formule classique de l'entropie de Shannon. Cette mesure, que nous appelons entropie nucléaire, ne tient alors pas compte de l'apparition de nouveaux mots, ni de la disparition de mots anciens, mais seulement de la diversité informationnelle du noyau résilient.

En plus de posséder l'avantage de permettre la comparaison entre différentes périodes, l'entropie nucléaire permet aussi de comparer plusieurs corpus. Il suffit simplement d'adapter la méthode en considérant l'intersection des noyaux résilients des corpus comparés à la place du noyau résilient initial.

9 Outils d'exploration

Dans ce chapitre, nous présentons des outils et interfaces permettant d'explorer les corpus de JDG et GDL sur la base des profils fréquentiels des n-grammes. Nous construisons le visualisateur de n-grammes pour ces corpus, munis d'options de recherches élargissant le champ des possibilités qu'offre habituellement cet outil. Nous présentons également une méthode permettant de créer une visualisation diachronique, le chronocloud, sur l'ensemble d'un corpus. Pour finir, une interface est également développée liant le chronocloud, le visualisateur de n-grammes et le site web de recherche dans les archives de JDG et GDL.

9.1 Visualisateur de n-grammes

Le visualisateur de n-grammes est un outil permettant d'afficher le profil fréquentiel d'un ou plusieurs n-grammes. Le visualisateur de Google qui permet de rechercher les profils fréquentiels des n-grammes jusqu'à n=5 sur des corpus de Google Books en plusieurs langues [1] est un exemple classique. Ce type d'interface permet à l'utilisateur d'effectuer une étude quantitative du comportement diachronique de certains n-grammes. On peut distinguer plusieurs cas de figure quant à l'utilisation d'un tel outil par un chercheur ou plus généralement un utilisateur. Dans un premier cas, l'utilisateur a une idée précise des n-grammes à étudier et peut tirer (éventuellement) des conclusions en fonction des profils fréquentiels affichés. Dans un deuxième cas, l'utilisateur recherche des n-grammes au comportement particulier (pics, tendance, apparition, disparition, cycles, etc...), utilise son intuition afin de cerner un groupe de n-grammes présentant un intérêt et peut tirer des conclusions quant aux phénomènes étudiés qu'il s'agisse de phénomènes historiques, linguistiques, culturels ou autres. Dans un troisième cas, l'utilisateur s'intéresse davantage à la relation observée entre les profils fréquentiels de certains n-grammes comme, par exemple, la recherche de ceux qui se font progressivement (ou même brutalement) remplacer par d'autres. Le chercheur utilise ces profils fréquentiels comme un faisceau d'indices afin d'appuyer ou infirmer une hypothèse formulée préalablement. Le visualisateur de Google est présenté dans la Figure 9.1.

1. Google n-gram viewer - https ://books.google.com/ngrams

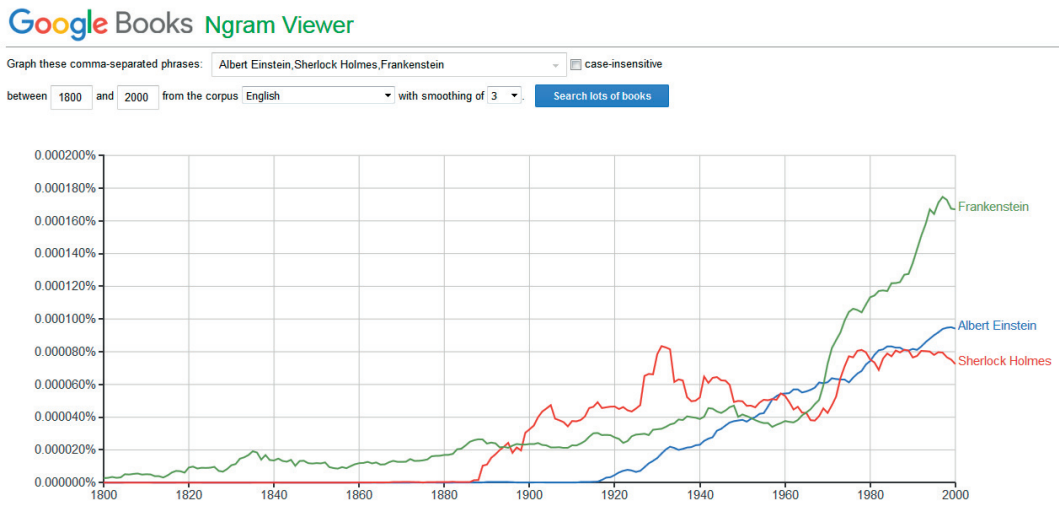


FIGURE 9.1 – Le visualisateur de n-grammes de Google avec son exemple par défaut, le 1-gramme "Frankenstein" et les 2-grammes "Albert Einstein" et "Sherlock Holmes" sur le corpus de Google Books en anglais de 1800 à 2000 et lissage par moyenne mobile de 3 années

La Figure 9.1 montre trois n-grammes dont les profils fréquentiels ont une tendance à l'augmentation monotone sauf quelques pics pour le 2-gramme "Sherlock Holmes". On remarque également que le 1-gramme "Frankenstein" est utilisé bien avant les 2-grammes "Albert Einstein" et "Sherlock Holmes". On observe que "Albert Einstein" et "Sherlock Holmes" ont été plus rapides dans leur augmentation de fréquence. Il est intéressant de constater que l'ordre en terme de fréquence de ces n-grammes change entre 1950 et 1970, période où ils ont une fréquence relativement équivalente. Etant donné que le lissage des courbes peut fausser leur interprétation, les prochains exemples seront montrés sans aucun lissage. Nous illustrons les profils fréquentiels de mots liés au thème du transport dans le corpus des Google Books francophones dans la Figure 9.2.

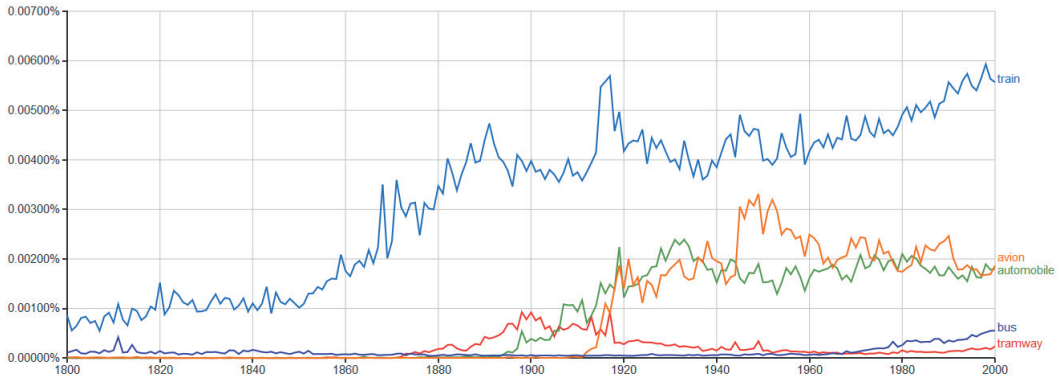


FIGURE 9.2 – Profils fréquentiels des 1-grammes "train", "avion", "automobile", "bus" et "tramway" dans le corpus de Google Books en français

Nous observons, dans la Figure 9.2, l'apparition de nouvelles dénominations de transports avec des fréquences différentes. Le mot "train" est le plus fréquent et aussi le plus ancien. Ensuite, apparaissent les mots "tramway", "automobile" et "avion". La fréquence du mot "bus" augmente lentement avec les années et celui-ci finit par être plus fréquent que "tramway" dans les dernières années avec un croisement de profils fréquentiels en 1966.

Afin d'étudier les corpus de JDG et GDL, nous avons construit un visualisateur de n-grammes dédié à ces corpus et pouvant en intégrer d'autres. Nous avons étendu le champ de possibilités d'étude de ces n-grammes en ajoutant diverses options de recherche.

year	gram	gram_1	gram_2	gram_3	gram_4	count	frequence
1 950	en ce qui concerne	en	ce	qui	concerne	760	0,0000965281
1 950	en ce sens qu	en	ce	sens	qu	20	0,00000254021
1 950	en ce sens que	en	ce	sens	que	35	0,00000444537
1 950	en cette fin d	en	cette	fin	d	16	0,00000203217
1 950	en cette fin de	en	cette	fin	de	13	0,00000165114
1 950	en chair et en	en	chair	et	en	11	0,00000139712
1 950	en chef de l	en	chef	de	l	32	0,00000406434
1 950	en chef de la	en	chef	de	la	29	0,00000368331
1 950	en chef des forces	en	chef	des	forces	33	0,00000419135
1 950	en collaboration avec la	en	collaboration	avec	la	12	0,00000152413
1 950	en collaboration avec le	en	collaboration	avec	le	14	0,00000177815
1 950	en collaboration avec les	en	collaboration	avec	les	19	0,0000024132
1 950	en collision avec un	en	collision	avec	un	13	0,00000165114
1 950	en collision avec une	en	collision	avec	une	18	0,00000228619
1 950	en compagnie de m	en	compagnie	de	m	13	0,00000165114
1 950	en compagnie de son	en	compagnie	de	son	12	0,00000152413
1 950	en contact avec la	en	contact	avec	la	13	0,00000165114
1 950	en contact avec le	en	contact	avec	le	11	0,00000139712
1 950	en contact avec les	en	contact	avec	les	13	0,00000165114
1 950	en corée du nord	en	corée	du	nord	38	0,00000482641
1 950	en corée du sud	en	corée	du	sud	34	0,00000431836
1 950	en couleurs parlé français	en	couleurs	parlé	français	27	0,00000342929
1 950	en cours d exercice	en	cours	d	exercice	19	0,0000024132
1 950	en cours d exécution	en	cours	d	exécution	13	0,00000165114
1 950	en cours de construction	en	cours	de	construction	15	0,00000190516
1 950	en cours de route	en	cours	de	route	27	0,00000342929
1 950	en d autres termes	en	d	autres	termes	68	0,00000863673
1 950	en dehors de l	en	dehors	de	l	24	0,00000304826
1 950	en dehors de la	en	dehors	de	la	28	0,0000035563
1 950	en dehors de toute	en	dehors	de	toute	16	0,00000203217
1 950	en demeure pas moins	en	demeure	pas	moins	18	0,00000228619
1 950	en direction de la	en	direction	de	la	25	0,00000317527
1 950	en droit d attendre	en	droit	d	attendre	14	0,00000177815
1 950	en droit de se	en	droit	de	se	15	0,00000190516

FIGURE 9.3 – Exemple de contenu de la base de données pour les 4-grammes du corpus JDG en 1950, commençant par le 1-gramme "en" et dont le nombre d'occurence est supérieur à 1

La première étape est de construire une base de données contenant tous les n-grammes jusque $n = 9$ ainsi que leurs fréquences absolues et relatives par année. Nous avons choisi le système de base de données MySQL afin de stocker ces informations. La valeur de la fréquence relative d'un n-gramme pour une année donnée est la fréquence absolue du n-gramme pour l'année considérée divisée par la somme des fréquences absolues de tous les n-grammes au cours de cette année. Ainsi, la somme des fréquences relatives de tous les n-grammes vaut 1

pour chaque année et cela les rend plus facilement comparables. Un index SQL est ensuite construit afin d'accéder plus rapidement à chaque n-gramme grâce aux requêtes SQL par l'interface online. Un exemple de contenu de la base de données pour les 4-grammes est illustré dans la Figure 9.3. Une fois que la base de données est prête à être utilisée, nous avons développé un script en langage php qui permet d'afficher une page web capable d'effectuer directement des requêtes MySQL sur la base de données ciblée.

Nous avons ajouté au visualisateur de n-grammes les options suivantes :
- Possibilité de sélectionner le corpus JDG ou GDL seuls, mais aussi les deux corpus simultanément, permettant de comparer les profils fréquentiels des n-grammes entre eux et au travers de plusieurs corpus.
- Possibilité de choisir entre trois prétraitements différents afin de mesurer l'effet des prétraitements appliqués aux corpus lors de l'extraction des profils fréquentiels des n-grammes.
- Possibilité d'exécuter une expression régulière permettant de combiner les profils fréquentiels d'un nombre arbitrairement élevé de n-grammes du corpus sélectionné. Exemple : la recherche de la somme des profils fréquentiels de tous les n-grammes se terminant par "er" pour cibler le type le plus courant de verbes à l'infinitif.
- Possibilité d'exporter des données au format CSV afin de les étudier via d'autres outils ou de reproduire les graphes avec d'autres designs.
- Possibilité de cliquer sur la courbe des n-grammes afin de rechercher ces n-grammes directement dans le site web de l'archive cherchant simultanément le n-gramme et l'année correspondant au clic de l'utilisateur.

Nous présentons cet outil dans la Figure 9.4.

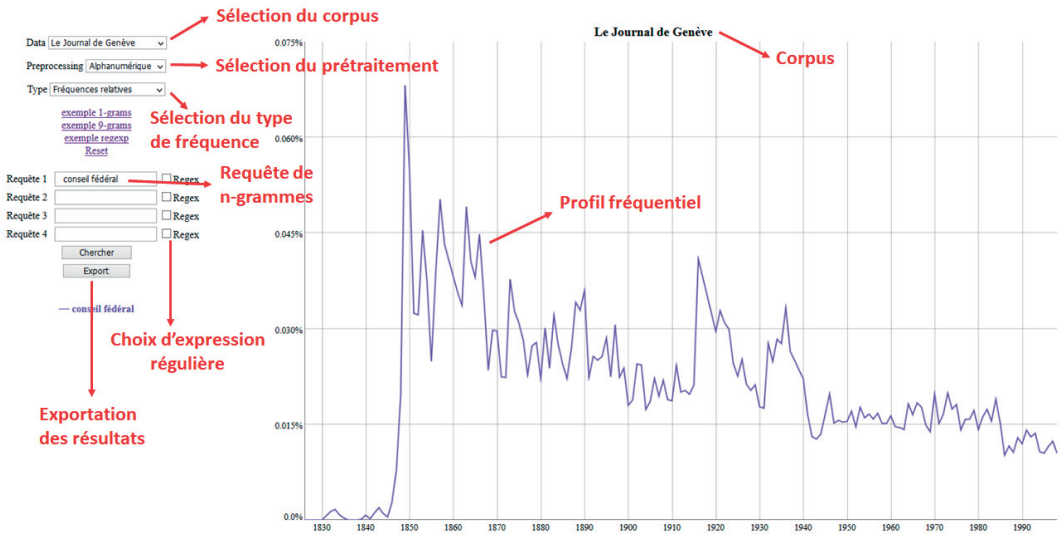


FIGURE 9.4 – Le visualisateur de n-grammes de JDG et GDL avec des annotations présentant l'utilisation de l'outil et les options disponibles sur l'exemple du 2-grammes "Conseil fédéral"

Les possibilités d'analyse des n-grammes à travers ce visualisateur sont nombreuses et variées. Considérons l'exemple d'une comparaison de mots entre "Russie" et "URSS" pour le corpus GDL, représenté dans la Figure 9.5.

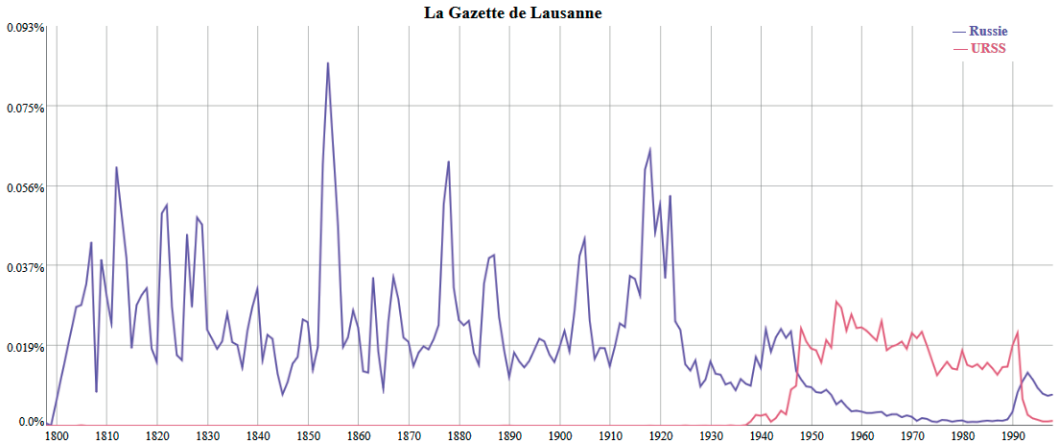


FIGURE 9.5 – Profil fréquentiel de "Russie" (mauve) et "URSS" (rouge) pour le corpus de GDL

Dans la Figure 9.5, nous observons un effet linguistique intéressant. Il semble en effet que les deux mots inversent leurs rôles en 1947 et que la courbe de "Russie" poursuit avec précision la courbe "URSS" jusqu'en 1992 où une seconde inversion de rôle est observée.

Les deux mots semblent être dans une situation de "lutte pour la survie" et nous observons l'émergence du mot "URSS", une stabilisation et une disparition aussi rapide que lorsque le mot est apparu. Les trajectoires des mots "URSS" et "U R S S" calculées dans les deux corpus sont présentées dans la Figure 9.6.

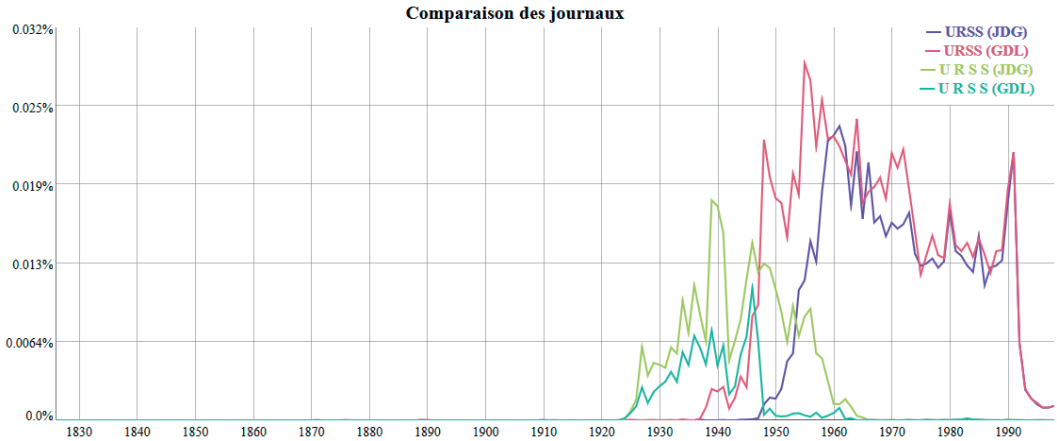


FIGURE 9.6 – Profils fréquentiels de "URSS" et "U R S S" pour les corpus de JDG (mauve) et GDL (rouge)

Dans la Figure 9.6, nous observons que l'apparition et la disparition de "URSS" sont attestés dans les deux journaux. Toutefois, la forme "U.R.S.S." fut utilisée avant "URSS", mais avec un décalage de 10 ans entre GDL et JDG. Cet exemple montre que les choix éditoriaux d'un journal se reflètent clairement dans certains profils fréquentiels. Cependant, même avec ce décalage, le même comportement global de "URSS" est observé et les profils fréquentiels de chaque journal finissent par se caler l'un sur l'autre. Nous allons revenir à ces exemples de différences de choix éditoriaux plus loin dans la thèse. Nous avons intégré une option "Regex" permettant à l'utilisateur de visualiser la somme des profils fréquentiels des n-grammes correspondant à une expression régulière. Par exemple, le chercheur peut vouloir trouver tous les 1-grammes qui représentent un nombre avec l'expression régulière simple "^[0-9]+$" identifiant tous les mots composés uniquement par des chiffres. Alternativement, on peut aussi vouloir rechercher tous les mots qui ne contiennent aucun chiffre via l'expression régulière "^[^0-9]+$". Le résultat des deux requêtes est illustré dans la Figure 9.7.

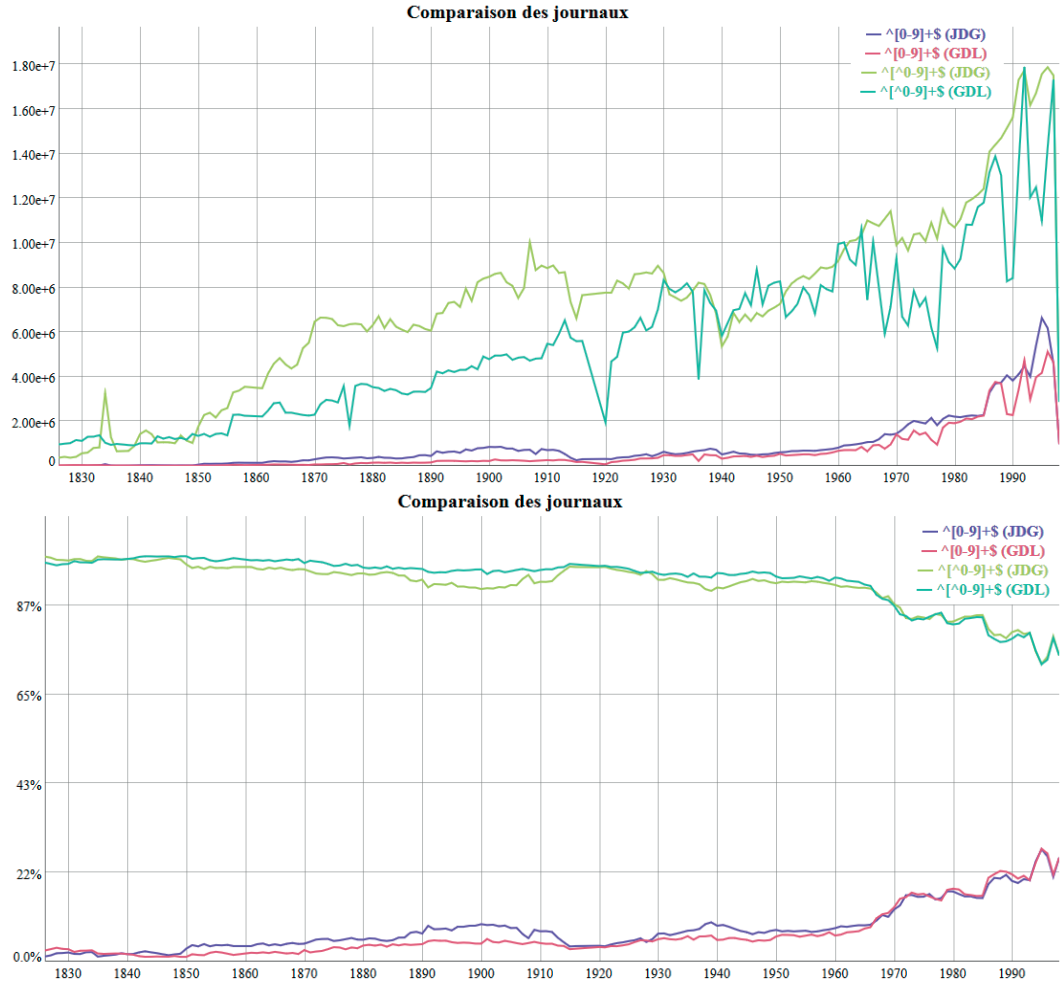


FIGURE 9.7 – Profils fréquentiels de mots correspondant aux expressions régulières "^[0-9]+$" et "^[^0-9]+$" pour les corpus de JDG et GDL en fréquence absolue (haut) et relative (bas).

Nous observons également sur la Figure 9.7, l'évolution croissante de la taille totale des données pour les deux journaux au travers des fréquences absolues. De plus, nous observons également l'augmentation parallèle des mots composés uniquement par des chiffres. La proportion des nombres augmente particulièrement après 1970, atteignant le taux de 25% pour les années les plus récentes. Un exemple de trois pages à forte représentation de nombres est donné dans la Figure 9.8.

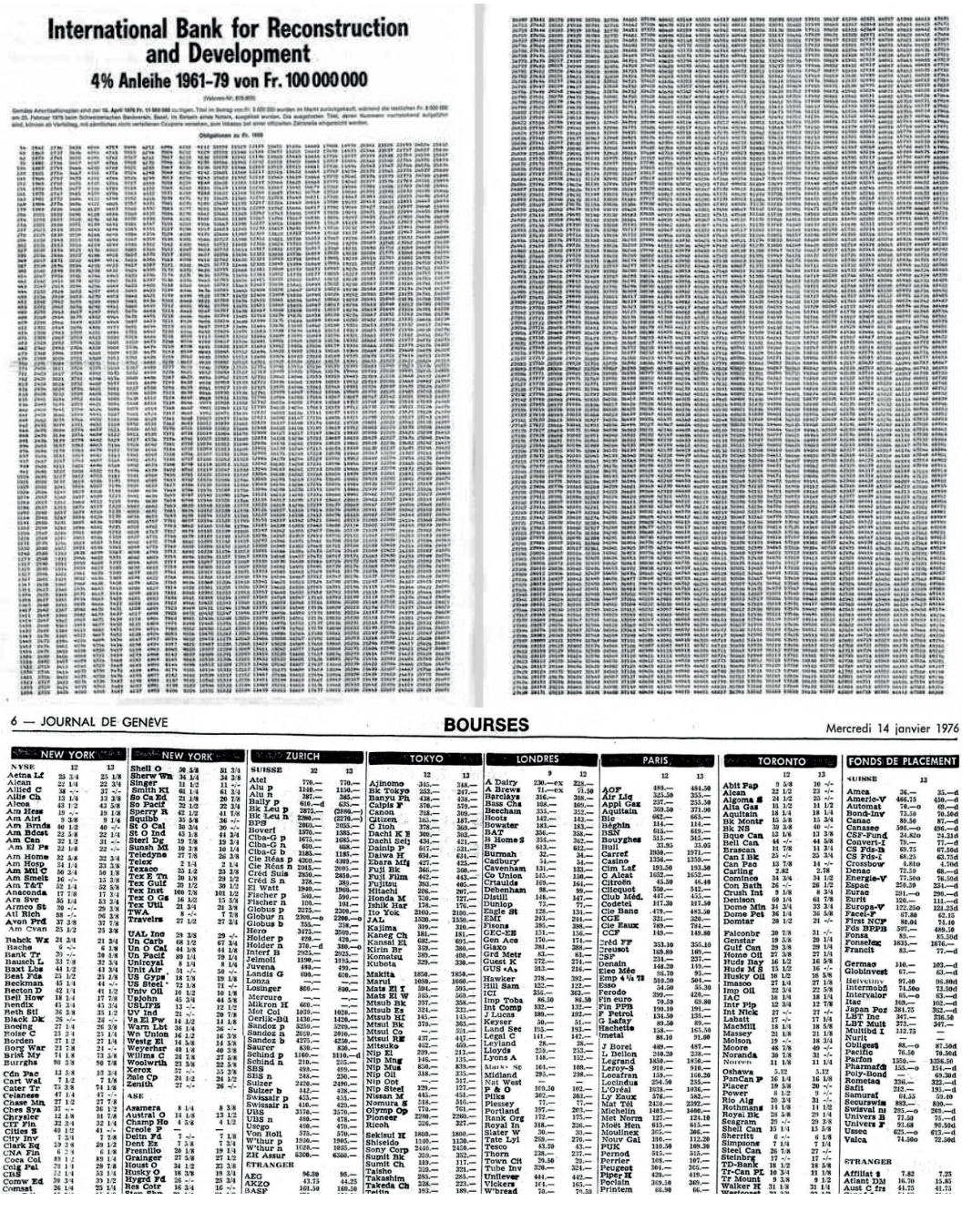


FIGURE 9.8 – Exemple de pages d'amortissement et de bourse dans JDG

Nous avons constaté que la principale raison de cette augmentation est l'introduction de nouvelles rubriques de journal sur les horaires de trains, de bus ou de tramways, les horaires de cinéma et en particulier des pages entières de chiffres, comme celles reprenant les valeurs d'amortissement en fonction de taux d'intérêt et la bourse, présentées dans la Figure 9.8.

Les nombres et leurs évolutions sont un sujet d'étude également intéressant ainsi que leurs relations avec les autres mots et n-grammes du corpus. Cependant, un certain nombre de profils fréquentiels de mots considérés comme fréquents peuvent être influencés par la fréquence relative des nombres composant les corpus, en particulier pour les années après 1970. Chaque profil fréquentiel doit donc être analysé avec prudence et selon les options de prétraitement alphanumérique et alpha afin de garantir que les effets observés ne proviennent pas de l'augmentation de la fréquence relative des nombre au cours du temps.

Un autre exemple d'utilisation d'expressions régulières est donné pour la recherche de tous les mots finissant par "er" et "ir" qui contiennent un nombre important de verbes réguliers et irréguliers à l'infinitif. Nous montrons les résultats de ces requêtes pour différents prétraitements dans la Figure 9.9.

La Figure 9.9 montre que le profil fréquentiel d'un mot fréquent ou bien des profils fréquentiels agglomérés peuvent être affectés par la proportion des mots représentant des nombres et oblige à être prudent dans les discussions concernant l'évolution des n-grammes. C'est pour cela qu'a été introduit la possibilité de visualiser également les profils fréquentiels selon d'autres types de prétraitements comme le prétraitement "alpha" qui ignore les mots représentant des nombres.

En effet, la partie supérieure de la Figure 9.9, représentant le prétraitement "alphanumérique", semble suggérer que la somme des mots terminés par "ir" et "er" diminue au cours des dernières années, mais la partie inférieure, représentant le prétraitement "alpha", contredit cette hypothèse. Tous les exemples de cette thèse seront habituellement affichés avec le prétraitement "alphanumérique", mais les vérifications ont été effectuées en parallèle sur les données correspondant au prétraitement "alpha" afin d'assurer la pertinence de nos observations et hypothèses.

D'autres exemples présentés dans la Figure 9.10 illustrent les différentes possibilités de normaliser (ou non) les fréquences de n-grammes : fréquences absolues (nombres annuels de mots), fréquences relatives (profil fréquentiel) et fréquences comparatives. La dernière option est proposée afin de faciliter les comparaisons de fréquences de mots en normalisant les fréquences uniquement sur les n-grammes recherchés. Cette option présente l'avantage d'annuler l'effet de corrélation entre les profils fréquentiels en comparant les variations uniquement en proportion.

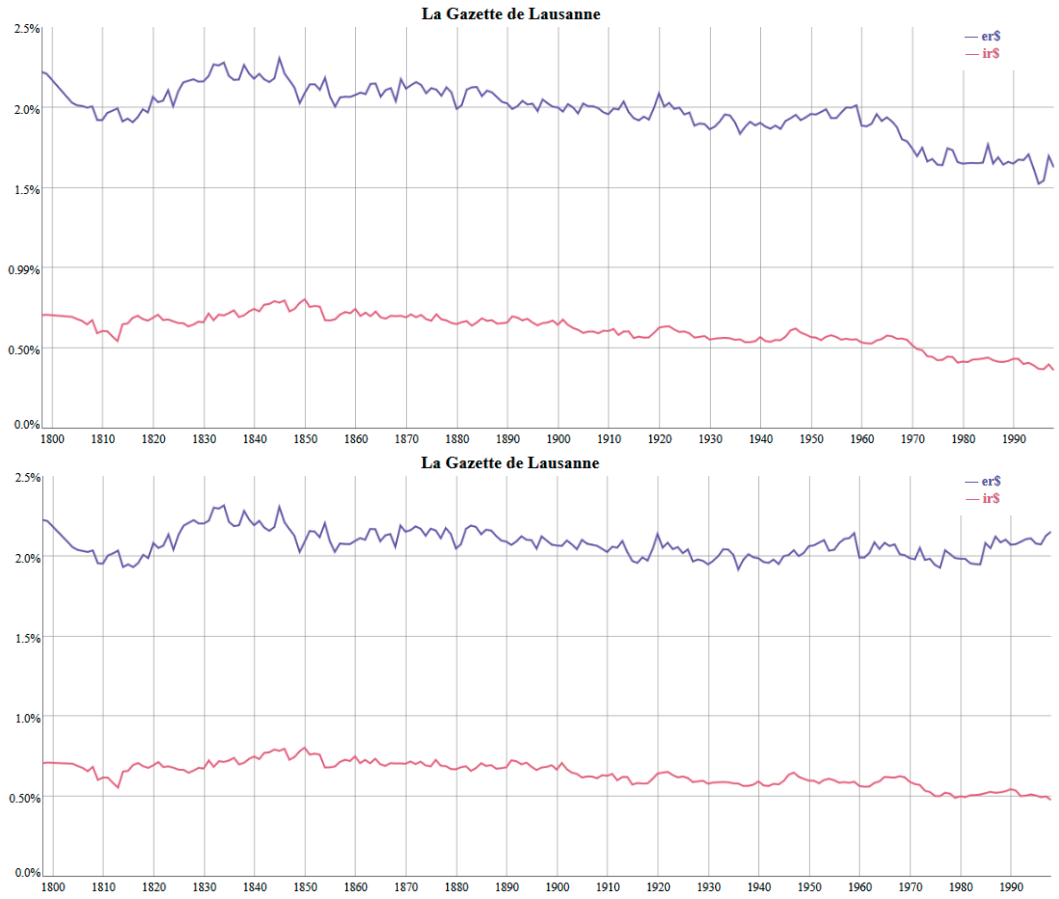


FIGURE 9.9 – Somme des profils fréquentiels des mots finissant par "er" et "ir" avec le prétraitement "alphanumérique" (haut) et "alpha" (bas)

Le visualisateur de n-grammes est un outil puissant qui permet de rechercher des profils fréquentiels de n-grammes particuliers donnant des indications sur les changements linguistiques, les tendances culturelles et l'évolution générale du corpus.

Une démonstration du potentiel des études basée sur la visualisation des profils fréquentiels est exposée dans le travail (Michel *et al.*, 2011) qui examine principalement les tendances culturelles à travers les corpus de Google Books. Le visualisateur de n-grammes présente l'avantage de fournir un large éventail de points de vue différents sur l'évolution du corpus selon les n-grammes recherchés.

Nous avons noté que la taille du corpus augmente de façon continue et que la fréquence des nombres augmente également avec les années. Nous avons observé par ailleurs plusieurs sources de perturbations pour l'analyse linguistique, en particulier dans les années après 1970 pour les deux journaux.

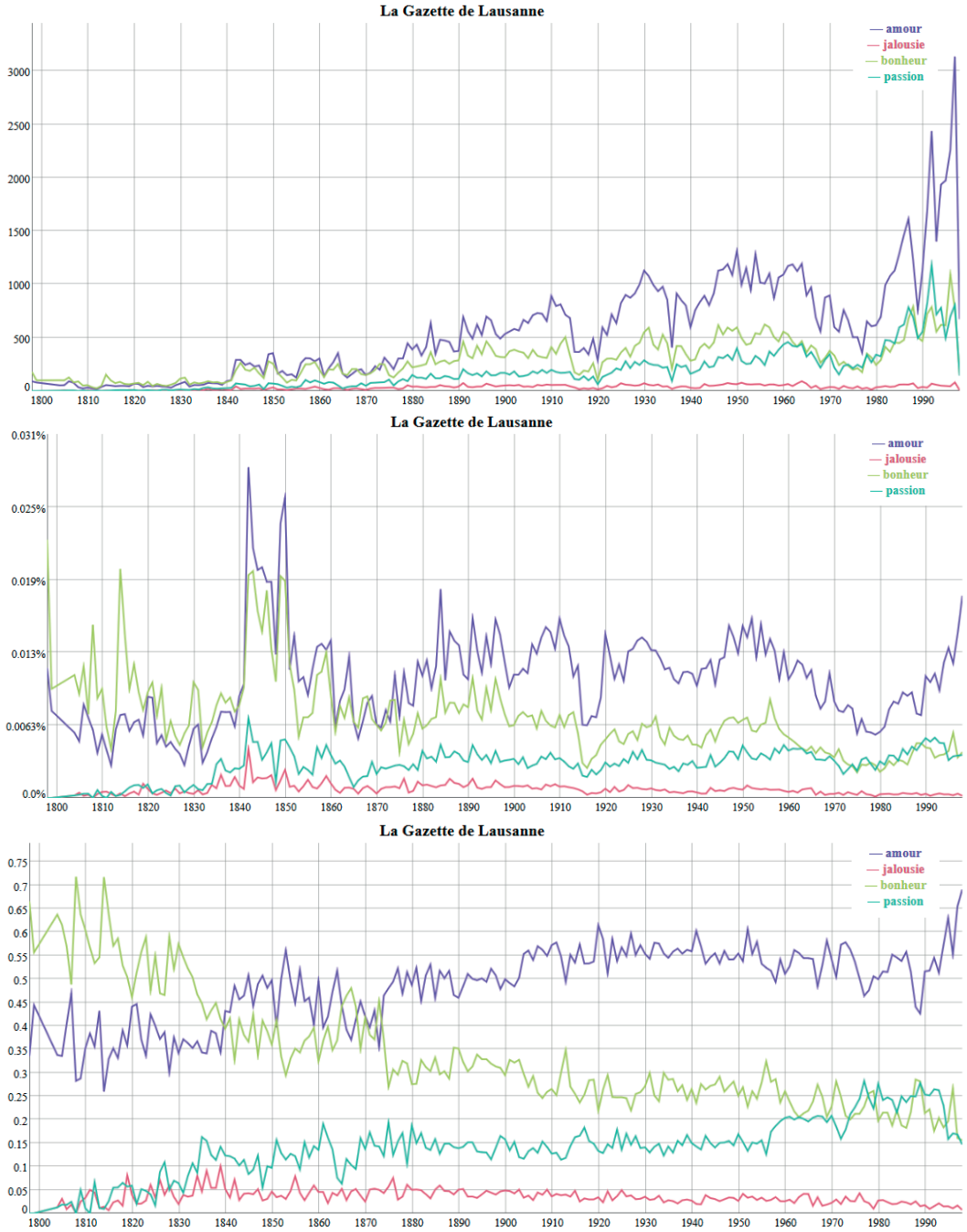


FIGURE 9.10 – Occurrences des mots "amour", "jalousie", "bonheur" et "passion" en fréquences absolues (haut), fréquences relatives (milieu) and fréquences comparatives (bas)

Les profils fréquentiels des n-grammes sont stockés dans des bases de données MySQL et sont interrogés en temps réel par l'outil en ligne, permettant également d'utiliser des expressions régulières, démultipliant les possibilités de recherche du visualisateur de n-grammes.

L'analyse d'un corpus à travers l'utilisation du visualisateur de n-grammes aide à appréhender son contenu et à caractériser les différentes évolutions linguistiques ou non linguistiques observées. Nous disposons également de la possibilité de rechercher des raisons ou des explications pour l'évolution des profils fréquentiels des mots grâce à une recherche directe sur le site des archives en un seul clic depuis le visualisateur de n-grammes.

Cependant, les profils fréquentiels des n-grammes ainsi que la recherche qualitative via la consultation directe des archives n'apporte pas de preuve formelle et mathématique de causalité, mais seulement une somme d'indices allant dans une même direction permettant de corroborer ou non les hypothèses culturelles et linguistiques faites sur l'évolution du corpus.

Le visualisateur de n-grammes peut être considéré comme un microscope agissant sur le corpus afin de trouver une évolution particulière et permet de regarder le contenu du corpus selon différents angles. Il caractérise ce que nous appelons le niveau Micro, un niveau d'étude de l'évolution des mots et des expressions individuelles au sein des corpus.

9.2 Chronocloud

9.2.1 Introduction

Dans cette section, nous présentons plusieurs concepts de visualisation permettant de représenter le contenu d'un corpus notamment au travers de combinaisons structurées de nuages de mots. Ces concepts ont également été présentés dans l'article (Buntinx *et al.*, 2017b). Nous y développons la visualisation chronocloud permettant de regarder un ensemble représentatif de mots et fréquences organisés autour d'un axe temporel.

Cette visualisation fournit donc une aide à l'étude de l'évolution du corpus et constitue un élément d'étude situé entre le niveau "micro" (niveau des n-grammes étudiés individuellement) et "macro" (niveau de l'ensemble du corpus). Le chronocloud est un outil d'exploration représentant les n-grammes et certaines de leurs caractéristiques dans un grand corpus de données temporelles. La visualisation chronocloud est composée de multiples nuages de mots (Hassan-Montero et Herrero-Solana, 2006) indépendants placés dans un espace de dimension 2, représentant plus d'informations sur les mots que la mesure des fréquences classiques. Cette visualisation est bien adaptée pour représenter la résilience des mots et le noyau résilient. Parmi les variantes possibles du chronocloud, nous présentons le chronocloud différentiel, développé afin de comparer deux corpus au sein d'une même visualisation.

Les nuages de mots se sont largement répandus et sont devenus très populaires ces dernières années (Viégas et Wattenberg, 2008), car ils permettent de connaître rapidement les mots les plus fréquents d'un corpus donné. Les principales raisons de leur popularité sont la simplicité d'interprétation et le design élégant (Seifert *et al.*, 2008) permettant aux utilisateurs d'apprendre des informations sur un contenu sans avoir à le lire.

Plusieurs études ont souligné comment les interfaces basées sur les nuages de mots (Sinclair et Cardew-Hall, 2008) peuvent aider les utilisateurs à naviguer et à trouver rapidement les informations souhaitées (Heimerl *et al.*, 2014), à prendre des décisions (Gottron, 2009), à explorer les entités nommées (Vuillemot *et al.*, 2009), à créer de nouvelles représentations graphiques (Riggs et Hu, 2013), à construire des outils pour la recherche (McNaught et Lam, 2010) ou pour étiqueter des documents (Seifert *et al.*, 2011). Les nuages de mots ont été utilisés afin de créer des représentations améliorées, montrant plus d'informations que la mesure classique de la fréquence globale des mots, comme le TreeCloud (Gambette et Véronis, 2010), SparkClouds (Lee *et al.*, 2010), Word Storms (Castellà et Sutton, 2013) (Castellà et Sutton, 2014) Ou RadCloud (Burch *et al.*, 2014).

Nous avons développé l'idée d'une nouvelle visualisation combinant plusieurs nuages de mots, organisés dans un plan muni d'un système de coordonnées polaires, permettant de représenter non seulement la fréquence des mots, mais aussi deux autres caractéristiques pour chaque mot. Nous utilisons les concepts de noyau et mots résilients développés précédemment et nous avons calculé les caractéristiques des mots en fonction de leur profil fréquentiel.

Nous déterminons la fréquence globale des mots, leur résilience dans le corpus ainsi que l'année correspondant à la fréquence maximale atteinte par le mot sur tout la durée couverte par le corpus. Nous visualisons le noyau résilient avec un nuage de mots qui a la forme d'un disque. Les mots qui sont suffisamment résilients pour apparaître dans la visualisation, mais pas assez pour apparaître dans le disque central font partie d'autres nuages de mots placés dans des couches externes dont l'éloignement du centre dépend de la résilience. En outre, le temps est représenté par la composante angulaire, mettant en évidence les mots qui font des apparences plus ponctuelles dans le corpus ainsi que le pic de fréquence de ces mots et nous pouvons donc visualiser une évolution temporelle du corpus dans un design de type montre. Seul le noyau résilient n'a pas de composante angulaire, car il est peu pertinent d'attribuer une année calculée à ces mots puisqu'ils se trouvent chaque année dans le corpus.

Les nuages de mots sont souvent utilisés pour visualiser rapidement les informations de contenu du corpus à un moment déterminé. Des travaux antérieurs ont essayé d'organiser les nuages de mots différemment afin de représenter une évolution du corpus (Collins *et al.*, 2009) (Weiwei *et al.*, 2010) (Coppersmith et Kelly, 2014). Nous avons défini précédemment (Buntinx *et al.*, 2016) (Buntinx *et al.*, 2017a) le concept du noyau résilient comme l'ensemble des mots les plus résilients dans le corpus, apparaissant au moins une fois par année couverte par le corpus. Ces mots sont toujours présents dans le corpus indépendamment de la période analysée. Un ensemble de mots de résilience R est défini comme un ensemble de mots qui sont présents au moins une fois par année sur une période maximale de R années consécutives. Un exemple visuel du calcul de la résilience et de l'année de fréquence maximale est présenté dans la Figure 9.11.

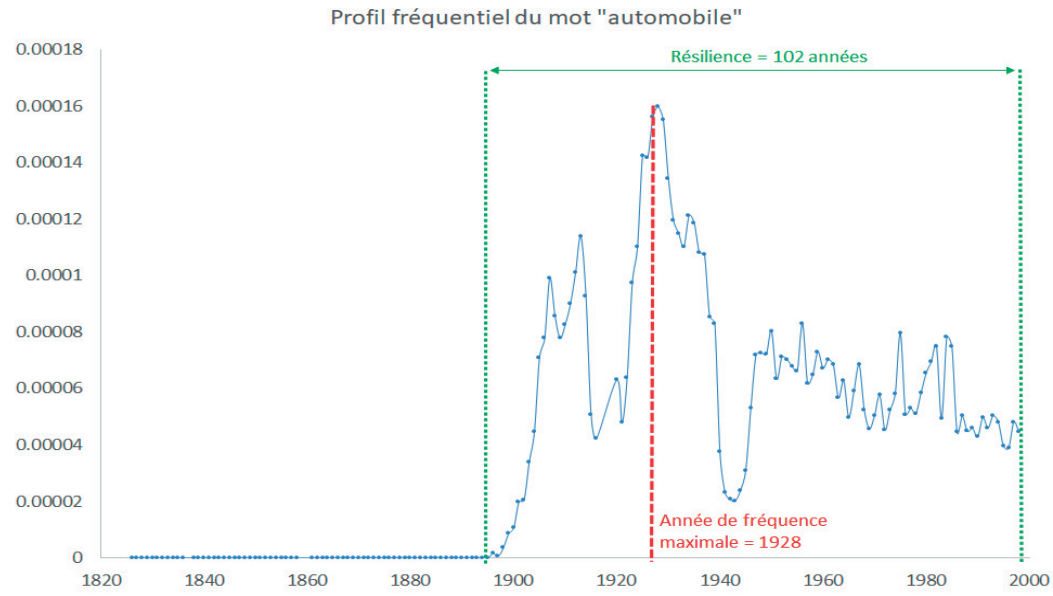


FIGURE 9.11 – Exemple de calcul de la résilience et de l'année de fréquence maximale du corpus avec le mot "automobile" dans le corpus JDG.

Sur un corpus d'une durée totale de 194 années (GDL) ou de 172 années (JDG), l'ensemble des mots de résilience 150 contient également le noyau résilient ($R = 194$ pour le noyau GDL et $R = 172$ pour JDG). Plus la résilience R augmente, plus les mots sont stables et plus l'ensemble de ces mots se rapproche du noyau résilient. Un nuage de mots affichant l'ensemble des mots de résilience 150 est représenté dans la Figure 9.12.

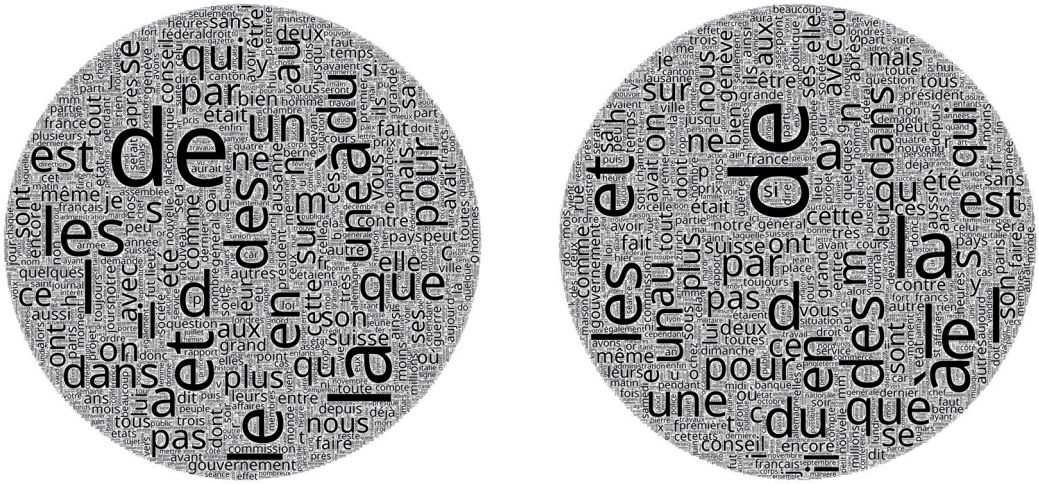


FIGURE 9.12 – Nuage de mots de l'ensemble des mots de résilience $R = 150$ pour GDL (gauche) et JDG (droite)

Dans la Figure 9.12, nous observons que les premiers mots que l'on visualise (donc ceux avec une police plus grande) sont principalement des mots fonctionnels (par exemple les articles, pronoms, conjonctions, adverbes et prépositions). Il n'est pas surprenant d'observer que ces mots sont parmi les plus fréquents, car ils sont utilisés dans un large éventail de contextes. Ces mots sont souvent considérés comme des "stop words" et retirés de la plupart des analyses, car ils sont trop communs et n'ont pas de relation spécifique avec un contexte particulier (Rajaraman et Ullman, 2011). Cependant, nous choisissons de garder tous les mots dans nos visualisations afin de ne pas faire d'hypothèses arbitraires sur les entités de base que sont les mots du corpus et de les traiter à égalité.

En zoomant sur la visualisation, comme sur la Figure 9.13, nous observons plus de mots spécifiques au contexte. Par exemple, nous identifions des caractéristiques géographiques comme "Suisse", "italienne", "autrichiens", "Angleterre" ou "Strasbourg" ainsi que des mots politiques et juridiques tels que "fédération", "annexion", "plaintes", "victimes", "occident" ou "réfugiés".

Afin de maintenir une représentation complète, avec un nombre important de mots résilients (dans un contexte d'analyse de grand corpus), le nuage de mots doit intégrer un zoom puissant permettant à l'utilisateur de visualiser tous les mots résilients présents dans le corpus. Les nuages de mots sont donc calculés en haute résolution et nous avons utilisé Deep Zoom

Images (DZI) avec OpenSeadragon [2], précalculant les fragments d'images pour chaque niveau de zoom avec MagickSlicer [3].

FIGURE 9.13 – Un exemple de zoom sur le nuage de mots de l'ensemble des mots de résilience $R = 150$ pour GDL

9.2.2 Polycloud : un espace structuré de nuages de mots

Bien que les mots résilients de 150 ans et plus soient déjà trop nombreux pour les représenter dans un nuage de mots sans possibilités de zoom, nous cherchons également à représenter des mots moins résilients, car ils contiennent aussi des informations intéressantes, en particulier sur l'évolution du temporelle du corpus. Par exemple, un mot qui existe au moins 50 années dans le corpus est aussi intéressant, car 50 ans de présence sont une indication de pertinence dans l'analyse de l'évolution et il est également cohérent d'étudier l'évolution de la fréquence (apparition, disparition, année de fréquence maximum) de ces mots. Le choix du seuil de résilience des mots à inclure dans la visualisation dépend du corpus et de sa couverture temporelle. De plus, le degré de résilience est une propriété des mots qu'on peut vouloir visualiser en parallèle à leur fréquence. Nous définissons ainsi le concept de polycloud.

Définition 14. *Un polycloud est une visualisation représentant une combinaison de nuages de mots structurée selon plusieurs caractéristiques dans un espace de dimension 2.*

Plusieurs algorithmes et idées d'implémentations et d'améliorations de nuages de mots existent (Kaser et Lemire, 2007) (Kaptein et Kamps, 2011) (Wu *et al.*, 2011). Nous avons utilisé un module Python [4] qui permet de combiner facilement les nuages de mots en une seule image. En raison de l'exigence de haute densité de mots, nous utilisons un ordinateur de type

2. OpenSeadragon - https ://openseadragon.github.io
3. MagickSlicer - Https ://github.com/VoidVolker/MagickSlicer
4. Python wordcloud - https ://github.com/amueller/word_cloud

cluster avec une grande mémoire (256 Go de RAM). Nous illustrons un polycloud composé de cinq nuages de mots dans une organisation en couches circulaires correspondant à différentes zones de résilience de mots. Le résultat est présenté dans la Figure 9.14.

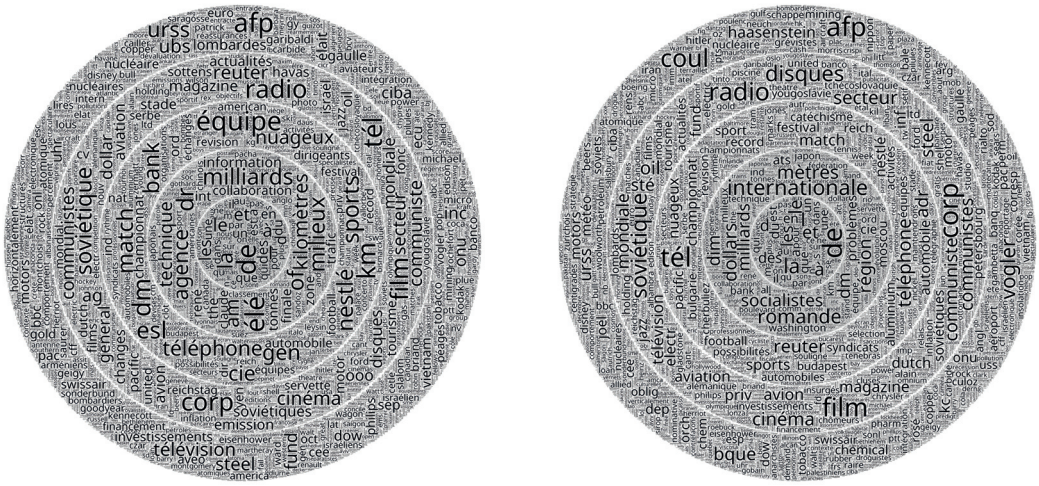


FIGURE 9.14 – Polycloud organisé en couches selon l'éloignement par rapport au centre ($R = 50-74$, $75-99$, $100-124$, $125-149$, 150 et plus, commençant respectivement de la couche la plus externe jusqu'au centre) pour GDL (gauche) et JDG (droite)

9.2.3 Chronocloud : illustrer l'évolution d'un corpus

Dans cette section, nous expliquons comment le chronocloud, cas particulier de polycloud, peut représenter l'évolution du corpus ainsi que les différentes couches d'ensembles résilients sous-jacentes au corpus. Nous introduisons une nouvelle dimension temporelle à savoir l'année où la fréquence des mots est maximale. Cette nouvelle caractéristique du mot est ajoutée via les coordonnées angulaires du plan en subdivisant le corpus en tranches de 10 ans, depuis 1840 jusqu'à 2000.

La visualisation de l'évolution temporelle suit donc le design d'une montre. Les mots résilients de 150 ans et plus ne sont pas subdivisés en catégories temporelles, car ces mots sont présents presque toutes les années dans le corpus. La couleur est une autre dimension qui a de multiples représentations possibles. Nous choisissons d'utiliser la couleur pour afficher la fréquence globale des mots, le bleu pour les basses fréquences, ensuite le vert, le jaune, puis le rouge pour les fréquences élevées.

Comme la fréquence est également une dimension illustrée par la taille des mots dans chaque nuage de mots, la couleur peut sembler à première vue superflue. Toutefois, étant donné que chaque nuage de mots est indépendant, la taille des mots ne peut servir qu'a comparer leurs fréquences au sein d'un même nuage. La couleur prend donc ici un rôle de liaison permettant de mettre en rapport les fréquences des mots sur l'ensemble de la visualisation.

Pour maximiser la lisibilité, nous utilisons une progression linéaire de la couleur avec un maximum déterminé. Nous avons choisi de définir ce maximum comme le minimum des 4 fréquences maximales de chaque quadrant, en considérant uniquement les mots qui ne font pas partie de l'ensemble le plus résilient ($R = 150$ et plus). En effet, ce calcul permet au chronocloud de montrer, au premier niveau de zoom, une diversité de couleurs et donc permettre une lecture plus naturelle et intuitive de l'évolution du corpus. Les chronoclouds de GDL et JDG en selon trois échelles de couleurs différentes sont présentés dans la Figure 9.15.

Les chronoclouds représentés dans la Figure 9.15 affichent trois caractéristiques de mots et peuvent potentiellement représenter (en considérant les possibilités de zoom) tous les mots qui se maintiennent au moins 50 ans dans le Corpus. Cependant, il est clair que le choix de l'échelle de résilience et du nombre de subdivisions temporelles dépendent des corpus étudiés et en particulier de leurs couvertures temporelles.

L'échelle de couleur [gris 50%,noir] semble mettre davantage en évidence les mots fréquents que les autres échelles. Cependant, il faut garder en tête que la variété des couleurs rencontrées va diminuer avec le niveau de zoom puisque nous avons choisi des paramètres optimisés pour le premier niveau de zoom. L'échelle de couleur [bleu,vert,jaune,rouge] est intéressante, car elle joue sur la sensation de couleurs froides à couleurs chaudes et possède un spectre plus large de coloris. Toutefois, certaines couleurs comme le jaune et le vert semblent moins visibles par rapport au bleu et au rouge. L'échelle [noir,rouge] permet de corriger ce défaut, mais réduit le spectre des coloris.

La dimension temporelle angulaire permet intuitivement d'observer une évolution temporelle du contenu du corpus, en particulier pour les mots qui apparaissent plus ponctuellement dans le corpus comme ceux de la dernière couche. En effet, la caractéristique de l'année de fréquence maximale des mots n'est qu'une projection d'informations sur un axe temporel et elle ignore d'autres années possibles de fréquence "élevée" des mots. Plus la résilience est élevée, plus cette projection risque d'entraîner une perte des informations sur l'évolution de la fréquence et, par conséquent, nous ne projetons pas les mots de résilience 150 sur la dimension temporelle.

Nous observons dans la Figure 9.15 que la partie noyau, au niveau de zoom initial (sans zoom), contient principalement des mots fonctionnels ou des stop words. D'autres couches contiennent des mots liés à des événements historiques, à des inventions humaines (radio, téléphone, télévision, etc.), à la géopolitique ou aux deux guerres mondiales.

Nous observons également, comme attendu, que plus les mots sont résilients, plus leur fréquence globale est élevée, toutefois ce n'est pas toujours le cas. Par exemple, le mot "URSS" a un profil fréquentiel similaire à un signal carré, commençant en 1944 et se terminant en 1996, mais avec des hautes fréquences. Il est donc très localisé dans le temps, mais à une fréquence globale élevée.

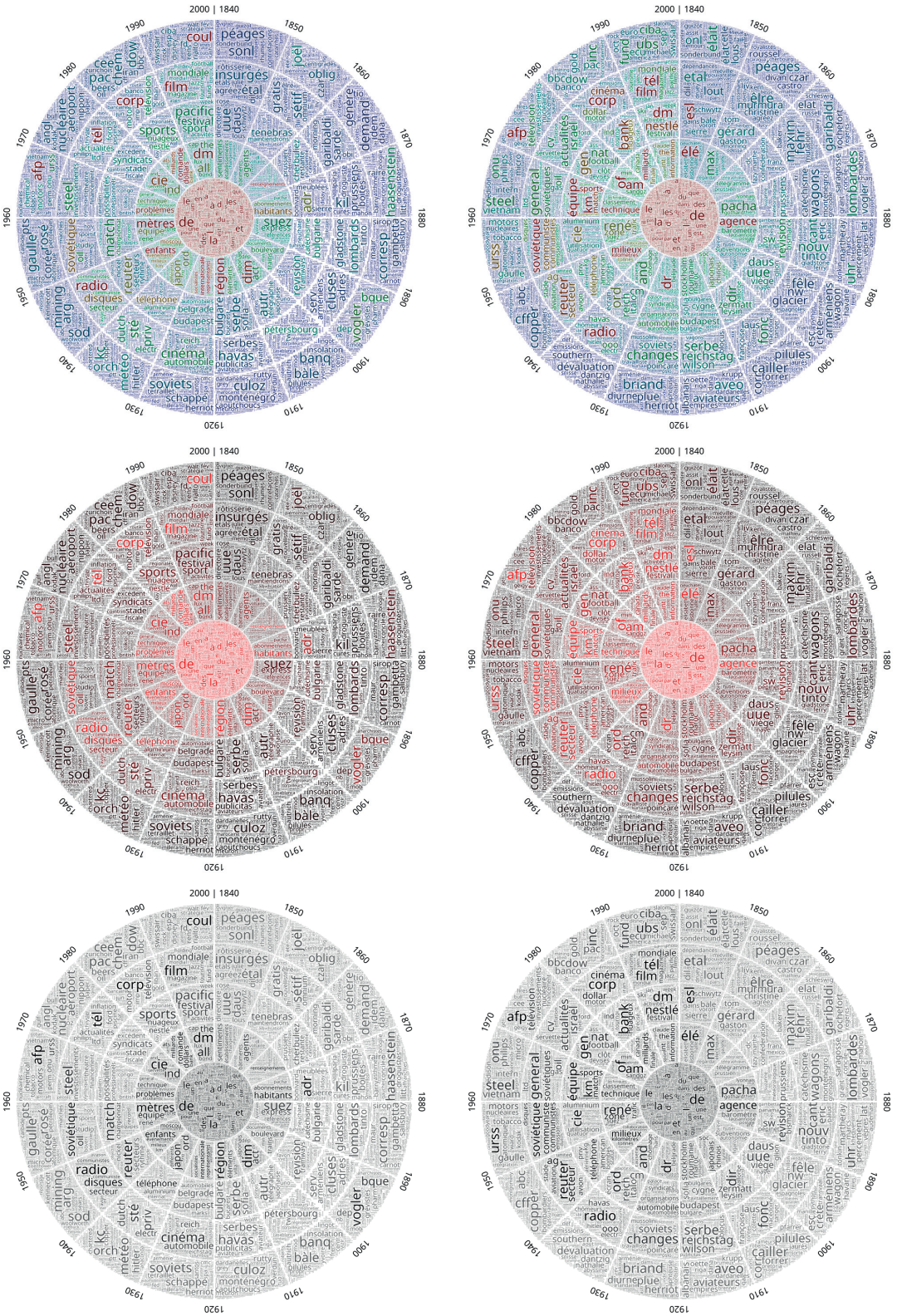


FIGURE 9.15 – Chronocloud pour JDG (gauche) et GDL (droite) en plusieurs échelles de couleurs, [bleu,vert,jaune,rouge] (haut), [noir,rouge] (milieu) et [gris 50%, noir] (bas)

9.2.4 Exemples de chronoclouds

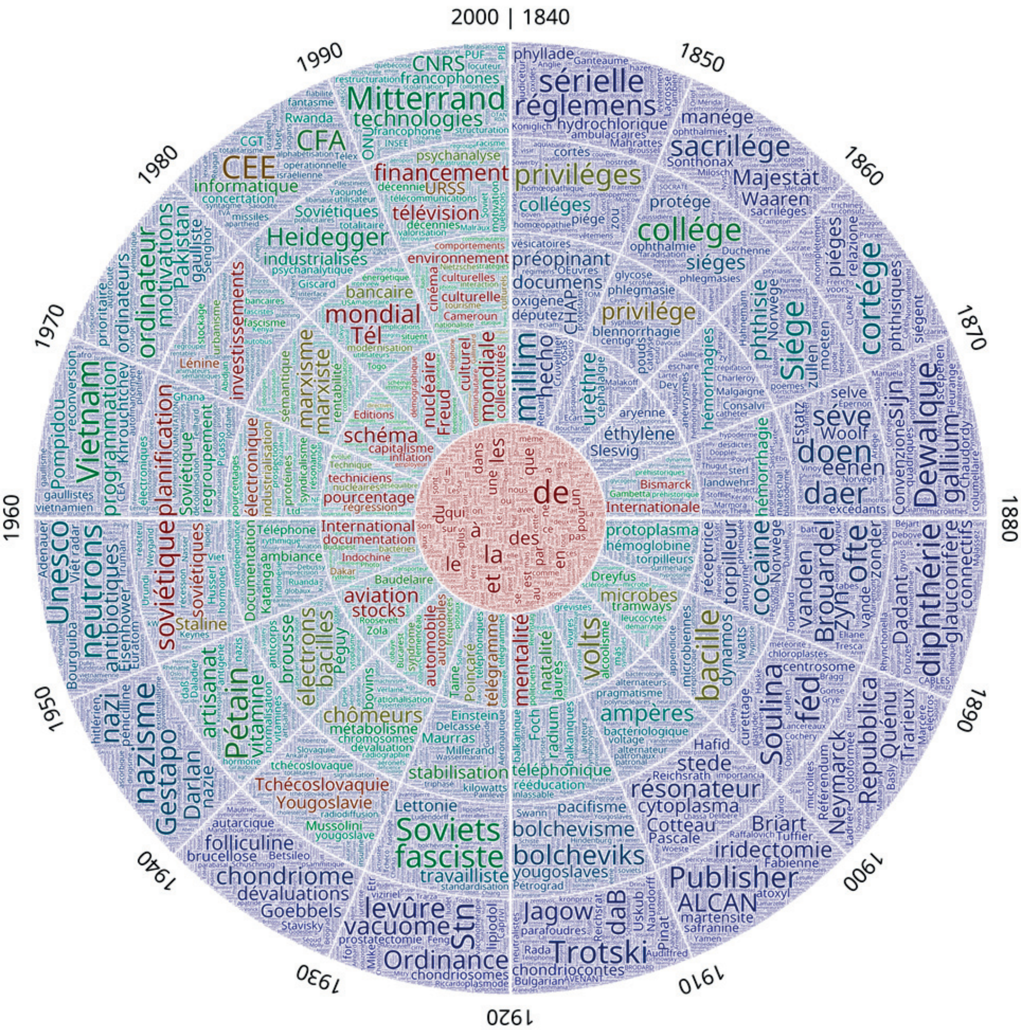


FIGURE 9.16 – Chronocloud pour le corpus de Google Books en français

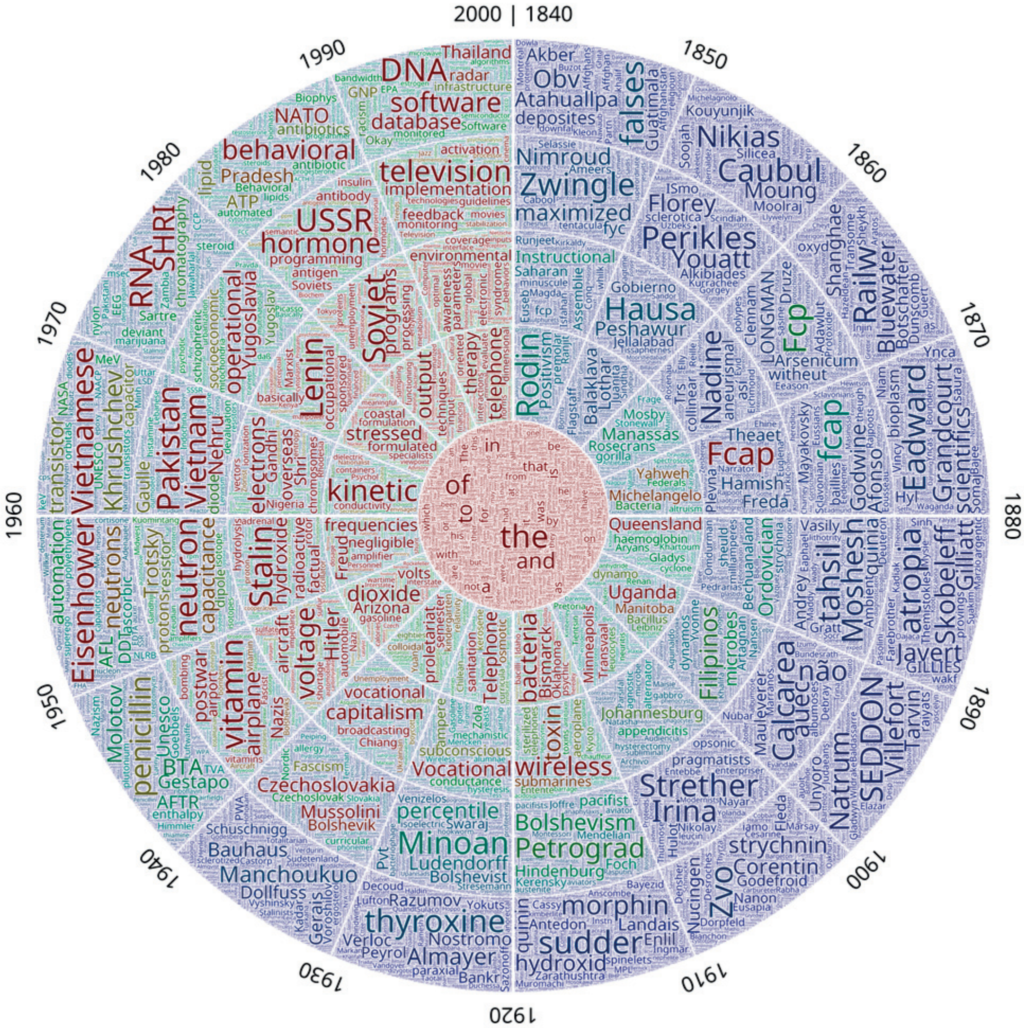


FIGURE 9.17 – Chronocloud pour le corpus de Google Books en anglais

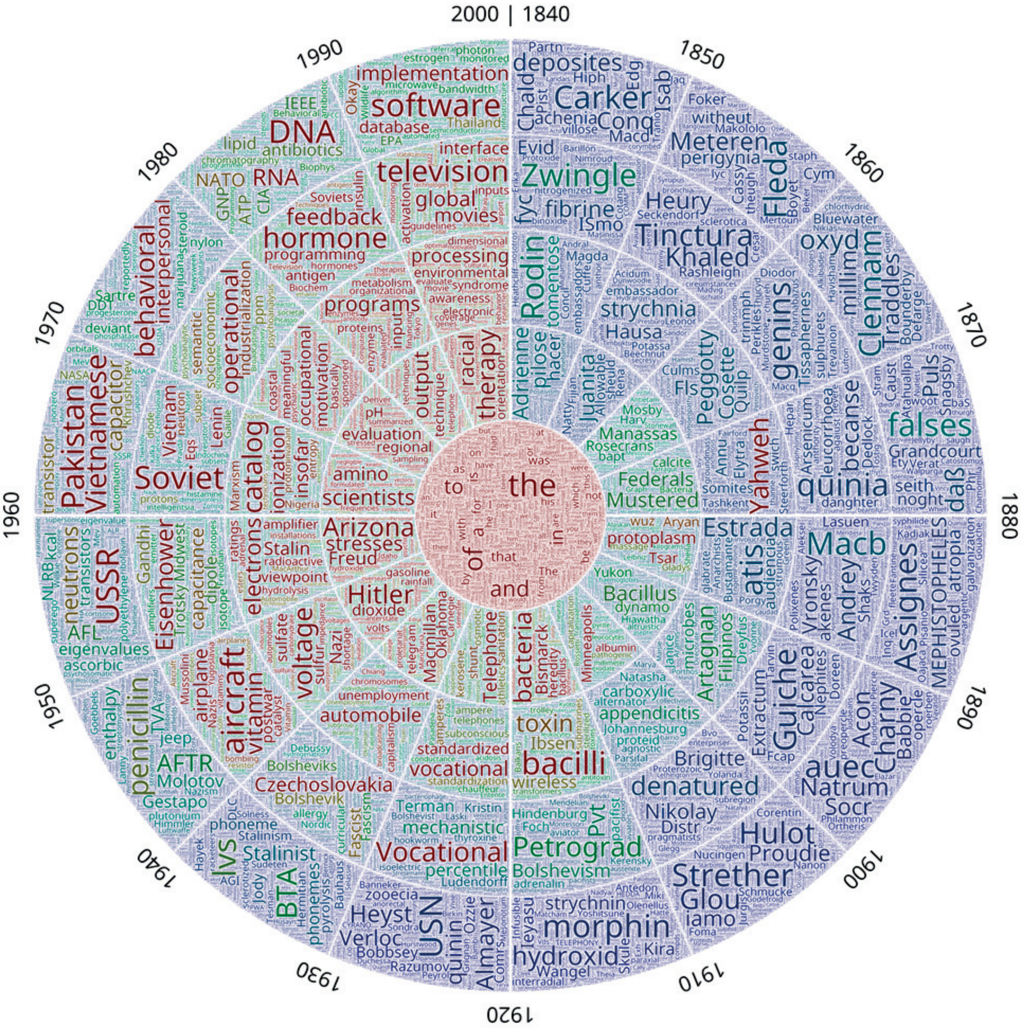


FIGURE 9.18 – Chronocloud pour le corpus de Google Books en anglais américain

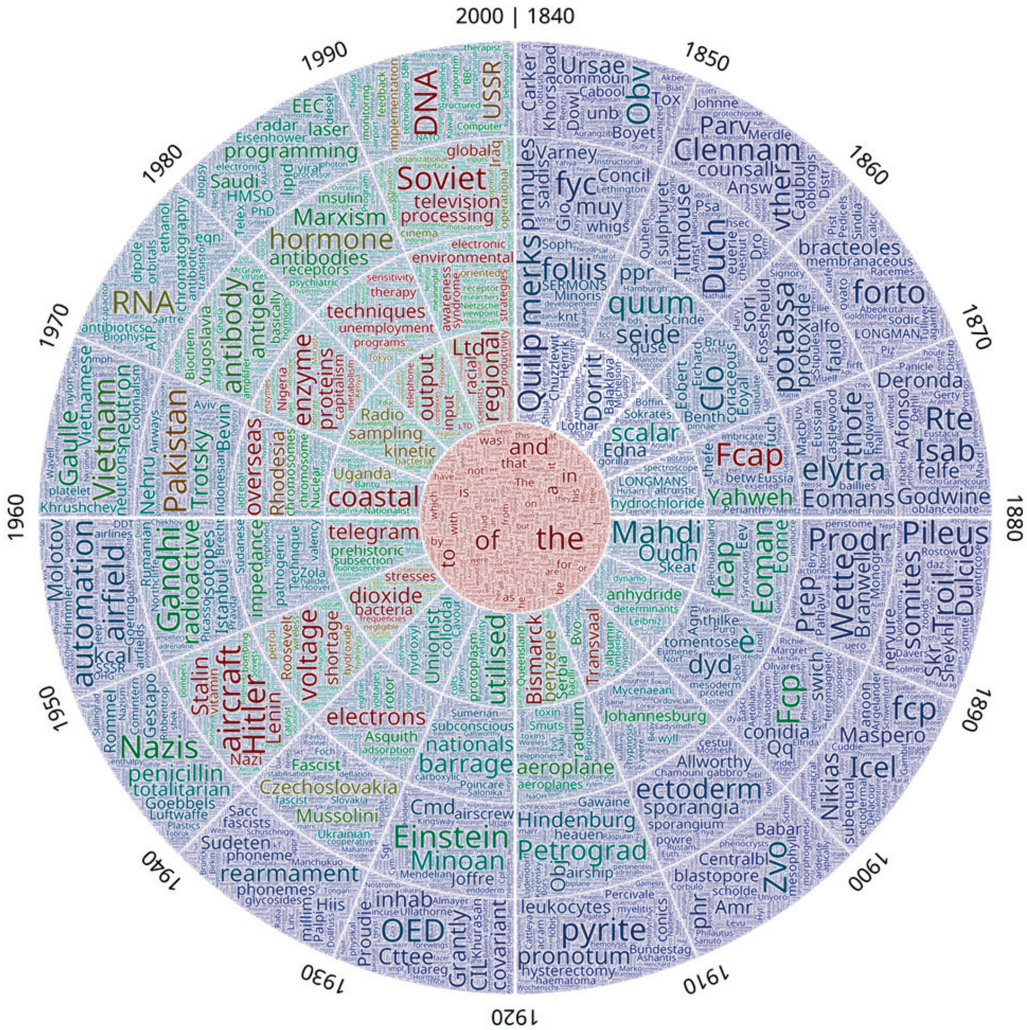


FIGURE 9.19 – Chronocloud pour le corpus de Google Books en anglais britannique

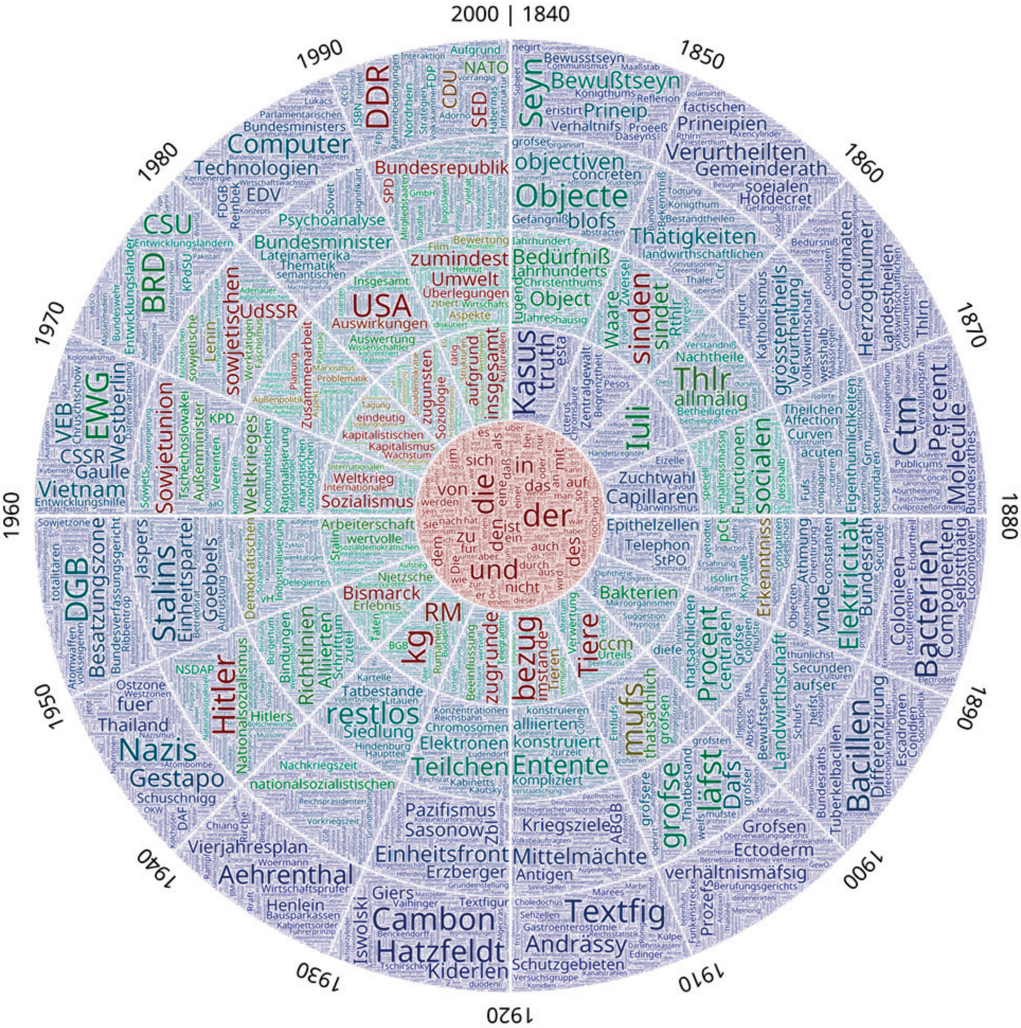


FIGURE 9.20 – Chronocloud pour le corpus de Google Books en allemand

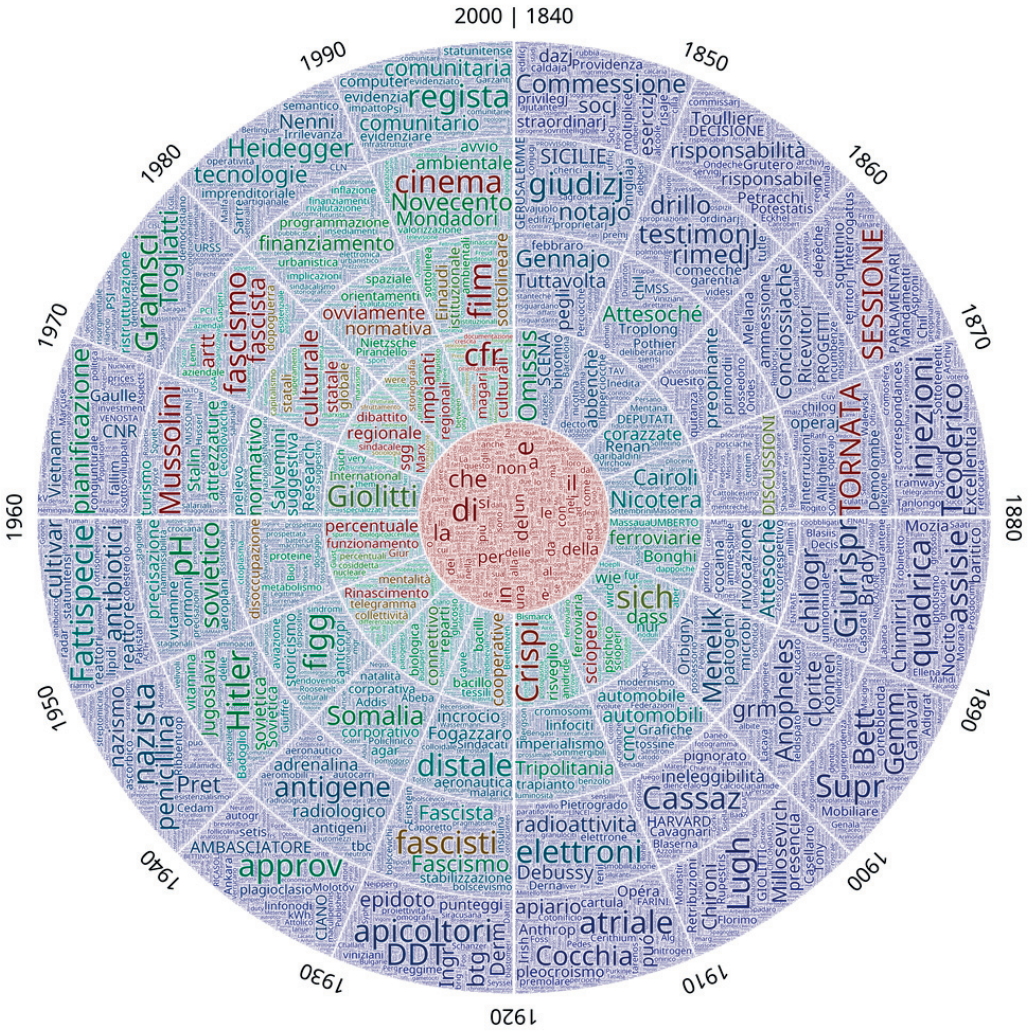


FIGURE 9.21 – Chronocloud pour le corpus de Google Books en italien

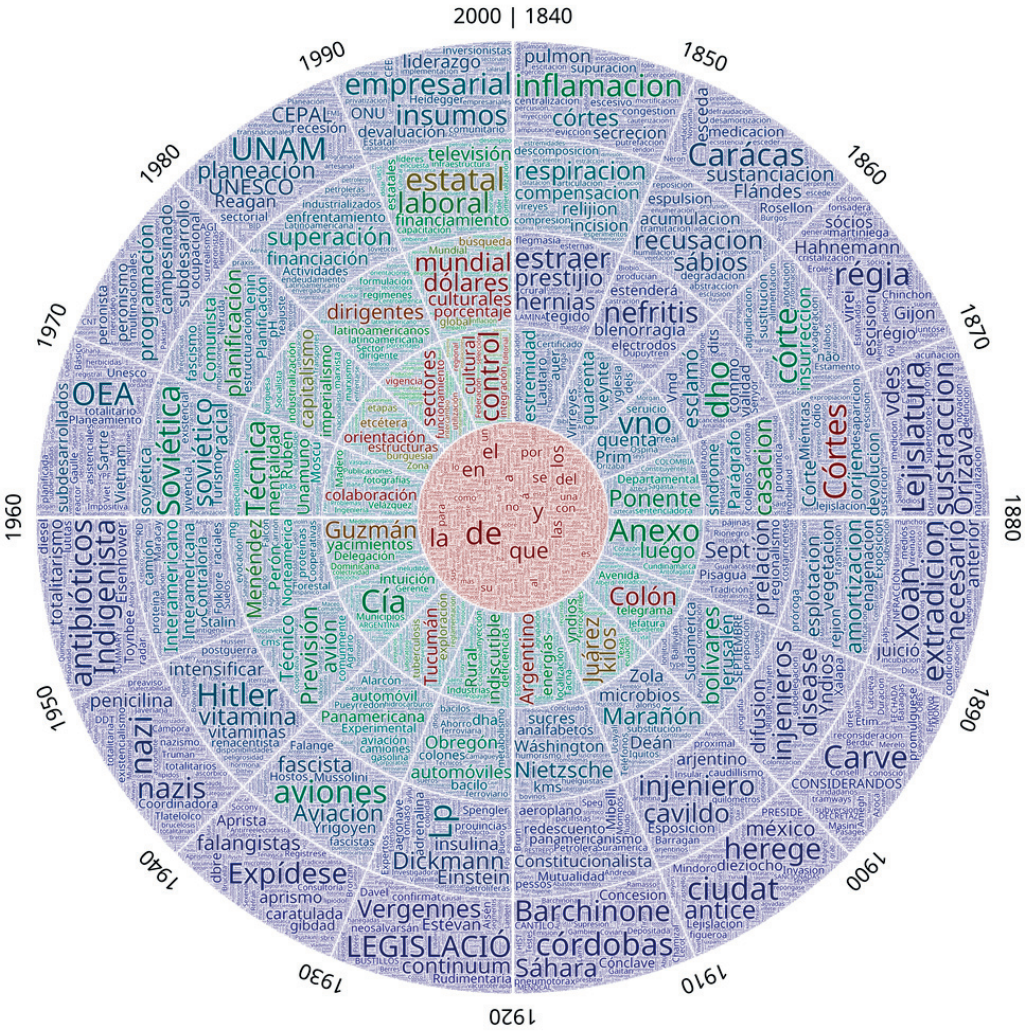


FIGURE 9.22 – Chronocloud pour le corpus de Google Books en espanol

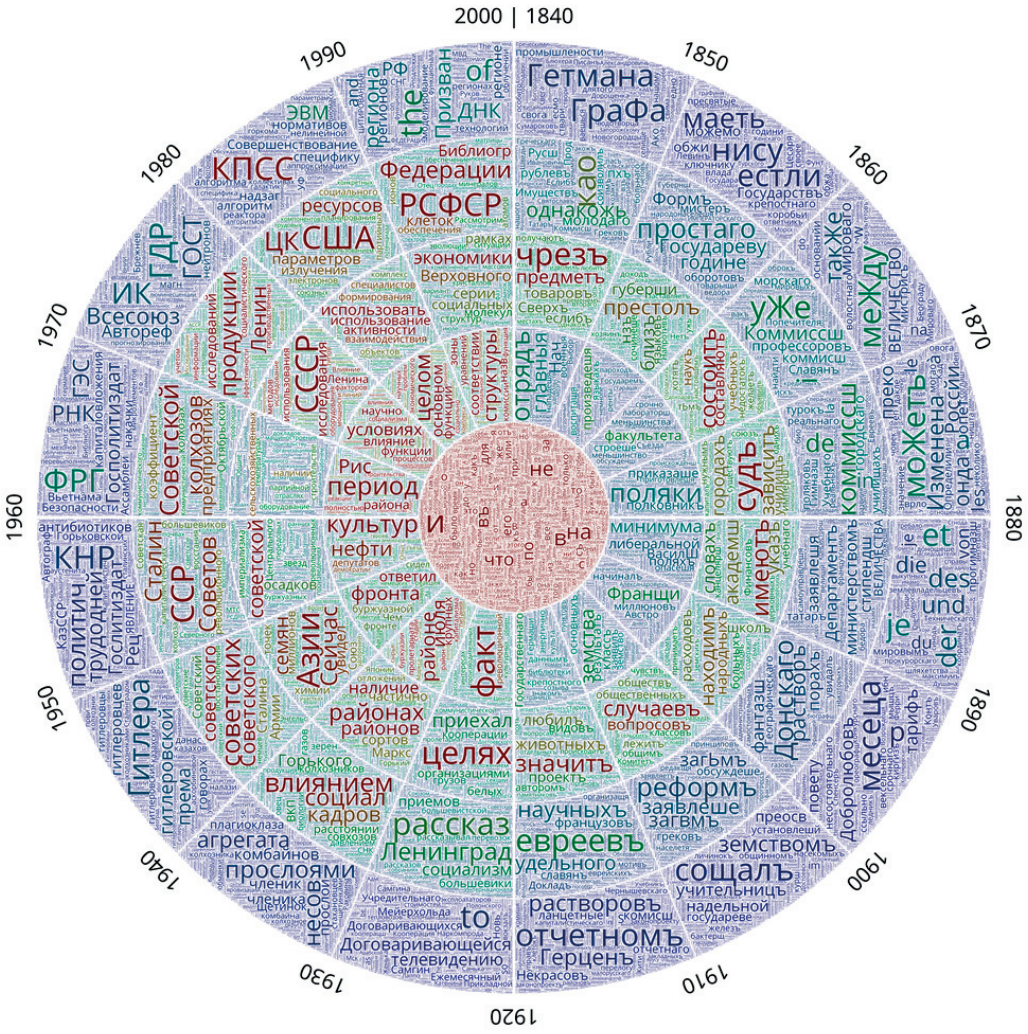


FIGURE 9.23 – Chronocloud pour le corpus de Google Books en russe

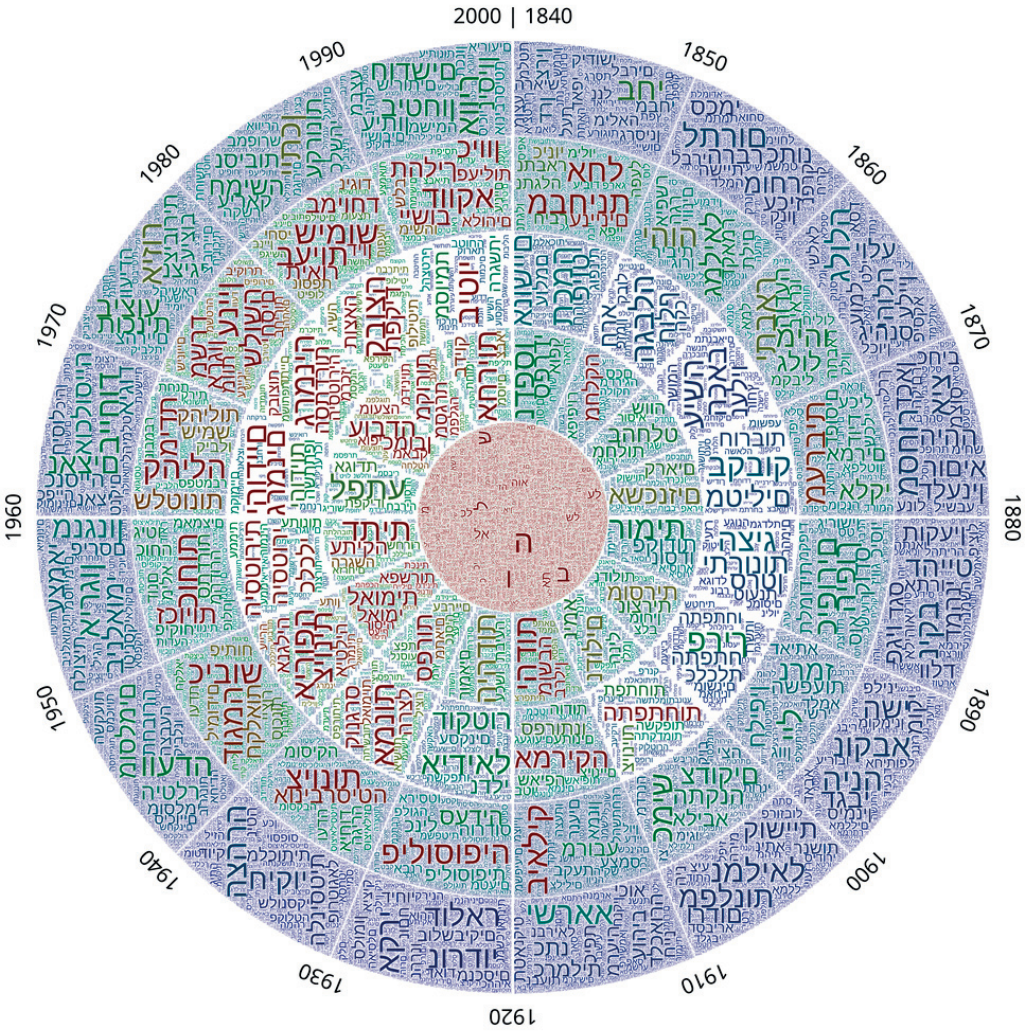


FIGURE 9.24 – Chronocloud pour le corpus de Google Books en hébreu

9.2.5 Chronocloud : un moteur d'exploration

La visualisation chronocloud aide les utilisateurs à voir les caractéristiques extraites des profils fréquentiels des mots comme leur fréquence globale, leur année de fréquence maximale et leur résilience dans le corpus. Nous avons amélioré le chronocloud avec une interface de type moteur de recherche permettant de mettre en évidence la position des mots, au niveau de zoom adéquat et dans une représentation de haute qualité (80'000 x 80'000 pixels). Le moteur de recherche [5] est présenté dans les Figures 9.25 et 9.26.

FIGURE 9.25 – Interface chronocloud en ligne avec possibilité de zoom profond

FIGURE 9.26 – Interface chronocloud en ligne affichant le résultat de la recherche du mot "industrie", révélant sa position et zoomant automatiquement au niveau adéquat

5. http ://chronocloud.epfl.ch

Pour les corpus de Google Books, nous avons relié l'interface avec le visualisateur de n-grammes de Google pour afficher le profil fréquentiel des mots. Pour les corpus GDL et JDG, nous avons relié l'interface avec le visualisateur de n-grammes expliqué dans la section précédente ainsi que dans l'article (Buntinx et Kaplan, 2015) et nous associons également celui-ci à l'interface de recherche du site web des archives de JDG et GDL (Rochat *et al.*, 2016). On peut donc cliquer sur les mots souhaités pour obtenir leur profil fréquentiel, comme le montre la Figure 9.27. Pour le corpus GDL et JDG, on peut cliquer sur le profil fréquentiel et se référer directement à la recherche du mot et de l'année ciblés dans les archives du journal, comme le montre la Figure 9.28.

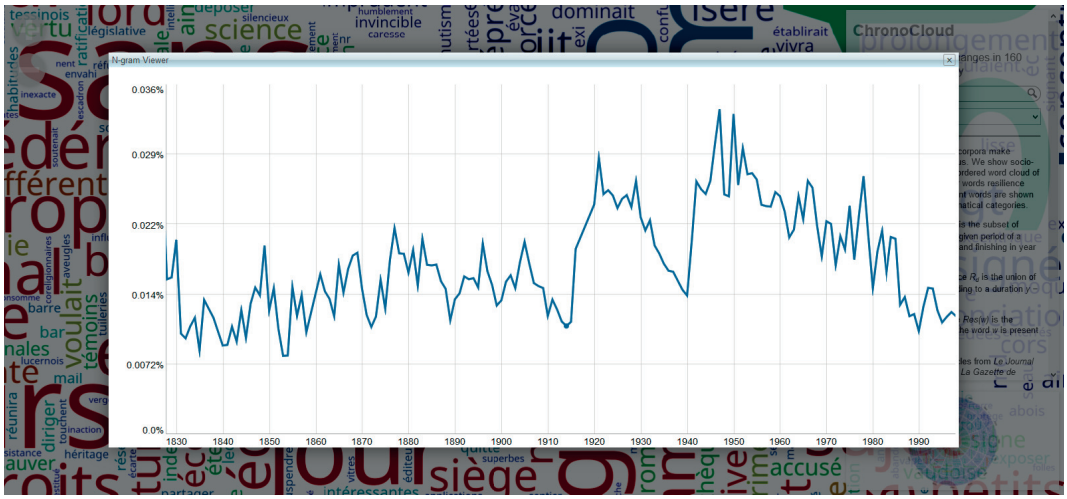


FIGURE 9.27 – Interface chronocloud en ligne affichant le profil fréquentiel du mot "industrie"

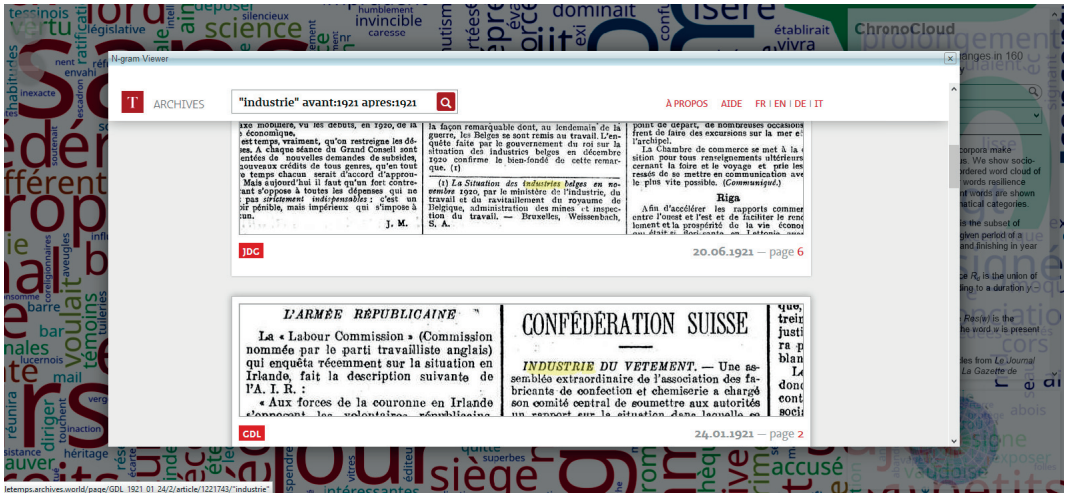


FIGURE 9.28 – Interface chronocloud en ligne et les résultats de recherche dans les archives pour le mot "industrie" en 1919

Il est important de noter que si le chronocloud peut révéler le contenu du corpus, les explications possibles de la mise en évidence des mots sur la base de leurs profils fréquentiels sont diverses. Il peut s'agir d'événements historiques, d'évolution linguistique, de pratiques éditoriales, de publicité ou même d'erreurs d'OCR. La recherche du type d'événement permettant d'expliquer certaines parties du profil fréquentiel observé peut généralement être traitée via l'exploration directe des archives.

9.2.6 Chronocloud différentiel

Le chronocloud donne une vision globale du contenu d'un grand corpus, mettant en évidence les différentes catégories de résilience, année de fréquence maximale et fréquences globales des mots. Cependant, il n'est pas aisé de comparer deux corpus via leur chronocloud. En effet, dans la configuration choisie, un mot de résilience 98 sera représenté dans le même nuage de mots qu'un mot de résilience 99. Cependant, un mot de résilience 100 sera représenté dans un autre nuage de mots, même si le mot de résilience 99 a exactement la même différence de résilience avec les deux autres mots. Cette propriété rend difficile la comparaison des mots selon les chronoclouds de deux corpus différents. Dans cette section, nous présentons le chronocloud différentiel, une variante de chronocloud appliqué sur deux corpus à la place d'un seul, révélant leurs différences. Ce chronocloud met en évidence des caractéristiques de mots calculées sur la base de profils fréquentiels différentiels entre deux corpus.

Le chronocloud différentiel représente les caractéristiques suivantes :
– La résilience des mots basée sur l'union des deux corpus.
– L'année où les profils fréquentiels différentiels des mots sont maximaux.
– Les fréquences moyennes des profils fréquentiels différentiels des mots.

Soit le profil fréquentiel G^1 du n-gramme g dans un corpus donné C_1 tandis que G^2 représente son profil fréquentiel dans le corpus C_2. Nous notons G_i^1 et G_i^2 la fréquence relative du n-gramme g respectivement dans les corpus C_1 et C_2 pour l'année i. Nous présentons deux variantes pour déterminer le profil fréquentiel différentiel ΔG permettant de définir deux types de chronocloud différentiel, l'un symétrique et l'autre asymétrique.

Soit deux corpus C_1 et C_2,

1. La variante symétrique du chronocloud différentiel est basée sur la définition de profil fréquentiel différentiel suivante :

$$\Delta G_i^{S_{12}} = \left| G_i^1 - G_i^2 \right| = \Delta G_i^{S_{21}}$$

2. La variante asymétrique du chronocloud différentiel est basée sur la définition de profil fréquentiel différentiel suivante :

$$\Delta G_i^{A_{12}} = max\left(0, G_i^1 - G_i^2\right) \neq \Delta G_i^{A_{21}}$$

Le chronocloud différentiel symétrique met en évidence les différences dans les deux corpus indépendamment de la "direction" de la variation, c'est-à-dire indépendamment du fait de savoir si la fréquence la plus grande est celle du premier corpus ou celle du deuxième. L'avantage de cette variante est de fournir une visualisation unique pour l'étude de la différence entre deux corpus. Le chronocloud différentiel asymétrique nécessite deux visualisations pour étudier les différences de corpus. Cependant, l'avantage de cette variante est de fournir les directions de variations, permettant de savoir, selon qu'un mot apparaît dans le chronocloud asymétrique ou dans le chronocloud asymetrique inverse, dans quel corpus sa fréquence est la plus grande. Cette visualisation sépare donc clairement la différence "positive" et "négative" entre les deux corpus et fournit ainsi une information plus complète sur l'analyse différentielle.

Le chronocloud différentiel ne peut être utilisé que pour comparer des corpus dont le lexique est proche. Par exemple, dans le cas d'une comparaison entre les corpus American English Google Books et British English Google Books, il peut révéler une différence linguistique dans l'utilisation de certains mots considérés comme des "américanismes". Les chronoclouds différentiels symétrique et asymétriques pour l'étude de la différence entre les corpus de American English Google Books et British English Google Books sont présentés dans les Figures 9.29. Le même type d'analyse peut être réalisé sur n'importe quelle langue moyennant un corpus assez large pour exploiter toutes les propriétés du chronocloud. La seule restriction pour le cas du chronocloud différentiel est que les corpus doivent partager suffisamment de mots communs.

Dans le chronocloud différentiel entre l'anglais britannique et l'anglais américain au travers des corpus de Google Books, nous observons des lieux géographiques, des personnalités, des termes politiques et ethniques. Toutefois, des particularités de l'anglais américain et de l'anglais britannique peuvent également être constatées au travers de certains mots possédant des différences marquées au niveau de leurs profils fréquentiels dans les deux corpus. Par exemple, les mots "aircraft" et "airplane" sont mis en évidence. Les fréquences de "airplane" sont faibles par rapport à celles du mot "aircraft" dans le corpus British English Google Books. Cependant, le mot "airplane", considéré comme un américanisme, a une fréquence non négligeable dans le corpus American English Google Books. Un autre exemple peut être observé sur le mot "gasoline" qui est typique de l'anglais américain et est utilisé à la place du mot "petrol" dans l'anglais britannique. Un raisonnement similaire peut être fait sur certains mots comme "automobile" et "car", "okay" et "OK" ou "organizational" et "organisational" qui ont une fréquence d'utilisation significativement différente dans les deux corpus.

Chronocloud permet aussi de visualiser les différentes caractéristiques des n-grammes avec $n > 1$. Ces chronoclouds ont la particularité intéressante de mettre en évidence les expressions multi-mots définies comme des interprétations idiosyncratiques qui traversent les limites fixées par les mots (Sag *et al.*, 2002)). Ces chronoclouds sont présentés et analysés pour n allant de 2 à 9 pour les corpus de GDL et JDG dans la partie suivante de la thèse.

9.2.7 Exemples de chronoclouds différentiels

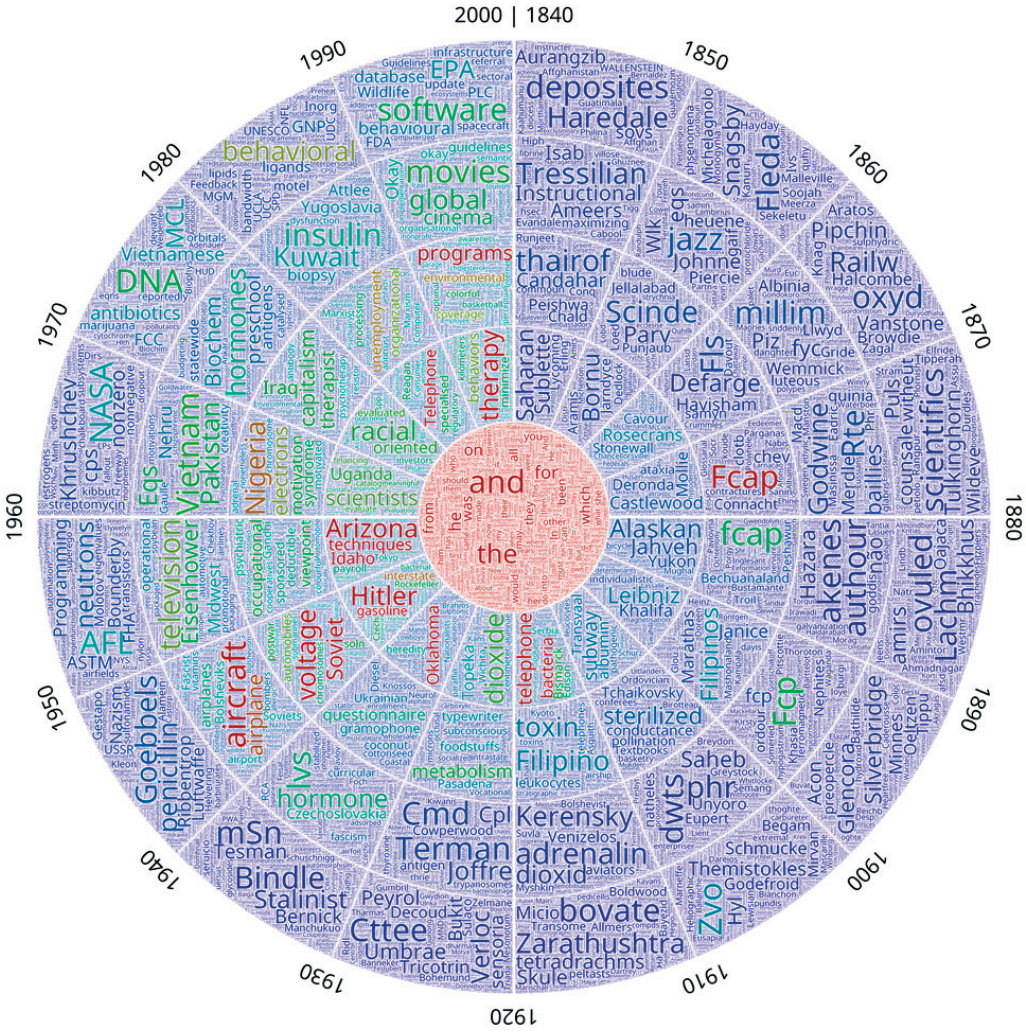


FIGURE 9.29 – Chronocloud différentiel symétrique pour les corpus American English Google Books moins British English Google Books

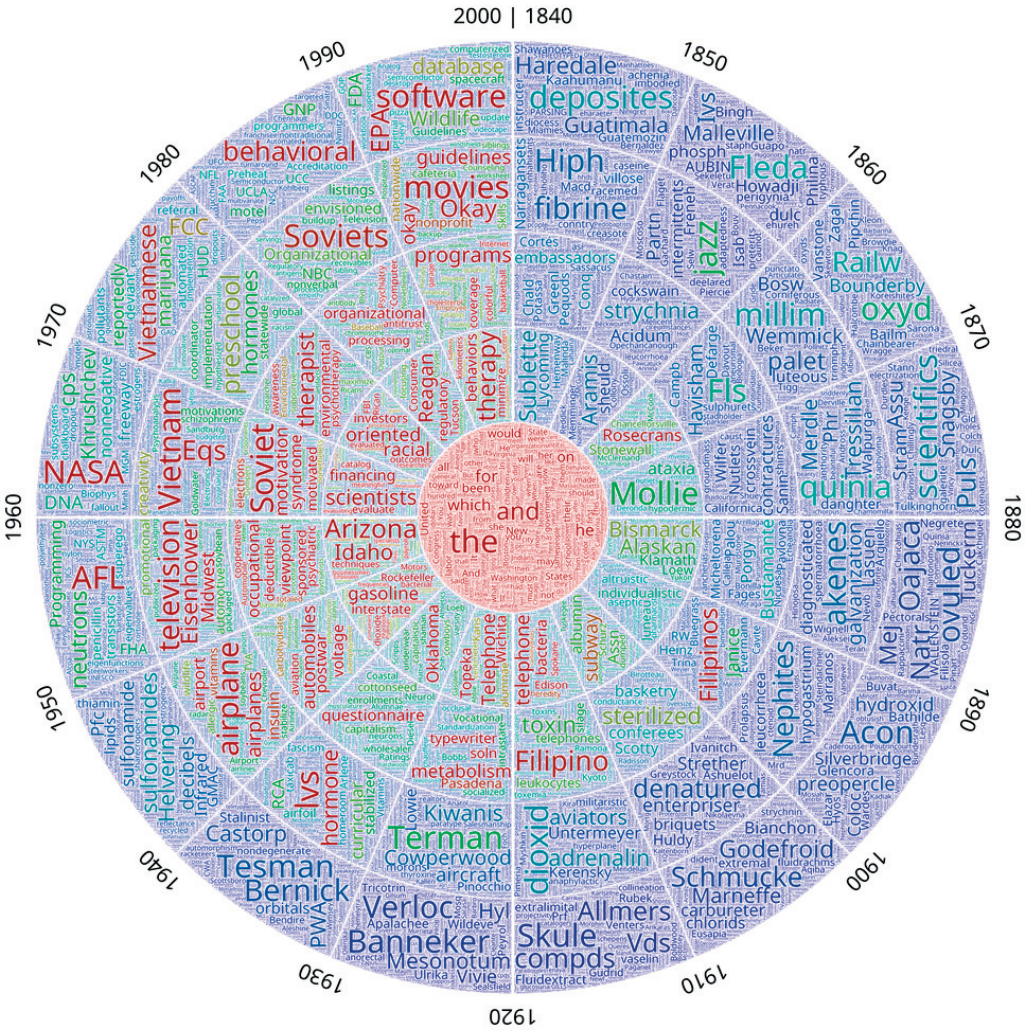


FIGURE 9.30 – Chronocloud différentiel asymétrique pour les corpus American English Google Books moins British English Google Books

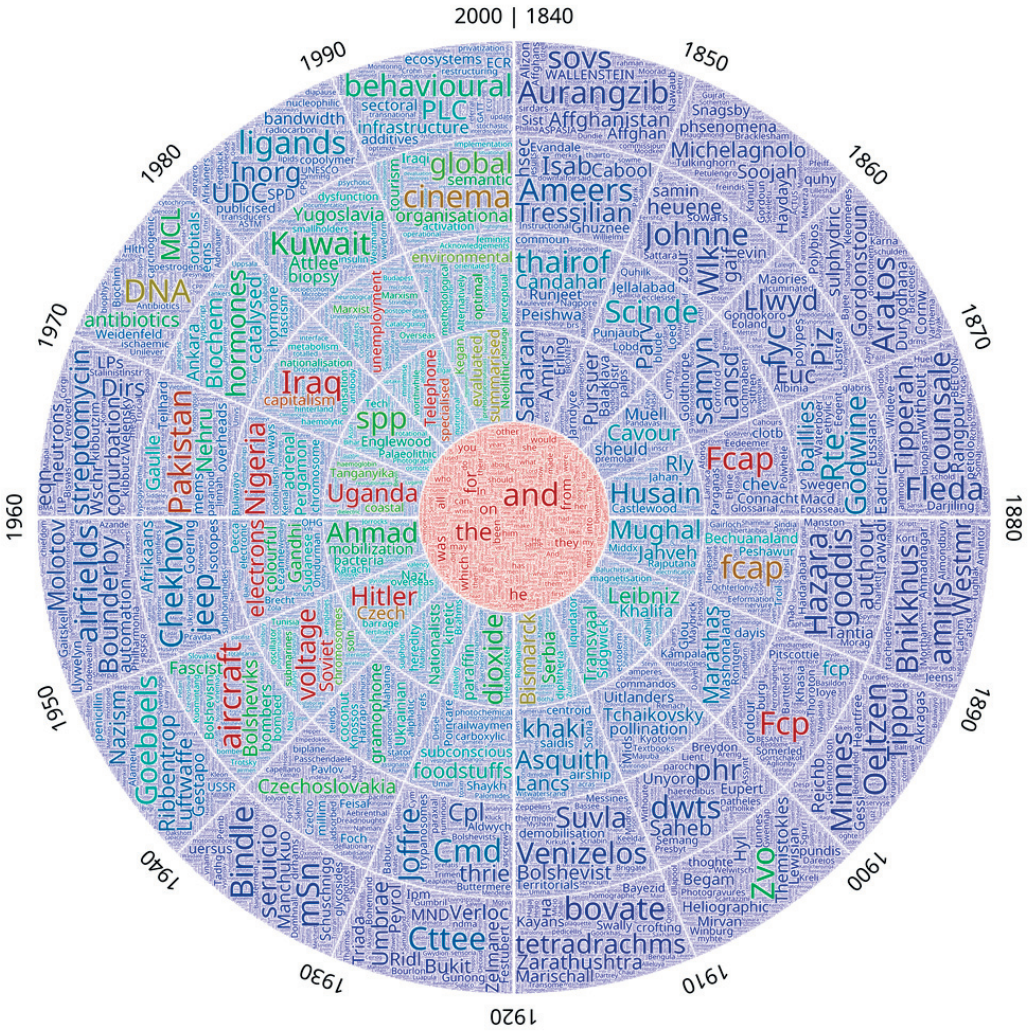


FIGURE 9.31 – Chronocloud différentiel asymétrique pour les corpus British English Google Books moins American English Google Books

En résumé, le chronocloud aide l'utilisateur à trouver des informations dans un corpus large et étendu dans le temps en se basant sur l'évolution des fréquences relatives annuelles des n-grammes. Ces informations permettent de naviguer et étudier le corpus pour en savoir plus sur des termes en particulier ou tout simplement pour étudier les raisons des pics de fréquence observés. On peut utiliser l'outil en ligne pour trouver des mots moins fréquents en faisant un zoom ou en recherchant directement le mot, pour ensuite naviguer à travers le visualisateur de n-grammes et finalement dans l'archive. Cet outil permet donc de traverser trois niveaux d'informations :

– Le niveau du chronocloud visualisant des informations sur tout le contenu du corpus en une seule visualisation.
– Le niveau du visualisateur de n-grammes affichant les profils fréquentiels des mots et aidant à sélectionner une période d'intérêt.
– Le niveau de la recherche directe dans les archives visualisant les journaux au sens propre selon les mots et la période souhaités.

Si le chronocloud a l'avantage de souligner les mots intéressants dans une perspective d'évolution temporelle, il faut noter que la détermination des premiers mots mis en évidence (sans zoom) peut avoir une sensibilité importante en fonction de la résilience de ces mots et de leur période de fréquence maximale. En effet, selon ces caractéristiques, nous classons les mots en plusieurs catégories. Par exemple, soit trois mots dont l'année de fréquence maximale est respectivement 1908, 1909 et 1910, le premier mot et le deuxième mot vont être classés dans la catégorie temporelle 1900-1909, mais le troisième sera classé dans la catégorie 1910-1919. Ce faisant, le deuxième mot qui pourtant est temporellement à la même distance du premier et troisième mot, sera classifié comme s'il était plus proche du premier mot. Ce problème de sensibilité généré par le modèle de découpage temporel est partiellement compensé par la possibilité de zoom et de recherche directe permettant de retrouver les mots concernés et observer leur profil fréquentiel. Toutefois, la navigation intuitive au travers du chronocloud selon les mots proches les un des autres restera impactée.

La visualisation d'un corpus de plusieurs millions d'articles dans une seule visualisation contenue dans un plan est un problème difficile et le chronocloud réussit à donner une visualisation structurée montrant des informations pertinentes sur le contenu du corpus. Les travaux futurs devraient explorer la possibilité d'une représentation plus robuste concernant les variations des caractéristiques en fonction des découpages pour la classification. Des variantes de chronocloud comme le chronocloud différentiel peuvent également être explorées afin de fournir de nouvelles façons de visualiser les différences de corpus. D'autres variantes peuvent être explorées pour améliorer la représentation des n-grammes avec $n > 1$.

En outre, l'exploration de nouvelles structures de polycloud ouvre de nombreuses possibilités de représentation du contenu des corpus et leur évolution, représentant les n-grammes et leurs caractéristiques outre la résilience, la date de la fréquence maximale ou la fréquence moyenne de ces n-grammes.

10 Aspects computationnels

10.1 Calcul distribué

La masse de données à traiter et à stocker impose notamment l'utilisation d'ordinateurs de type cluster avec des caractéristiques de processeur et mémoire suffisamment élevées. Nous avons utilisé un cluster qui dispose des caractéristiques suivantes :

- 2x12cores@2.5 GHz
- 256GB RAM
- 2x240GB SSD
- 6x4TB@7200RPM
- 2xGPU (NVIDIA TITAN X)

Outre l'infrastructure, il est utile de disposer de compétences dans le domaine du calcul distribué afin d'exploiter au maximum les capacités du cluster et améliorer les temps d'exécutions des différents algorithmes.

Les algorithmes développés dans le cadre de cette thèse utilisent presque exclusivement le langage de programmation Python qui est un langage de programmation répandu permettant une certaine agilité notamment parce qu'il dispose de nombreuses librairies complètes et bien documentées. En outre, c'est un language universel plutôt facile à apprendre et qui s'est propagé rapidement dans le monde de la recherche et particulièrement dans le domaine des humanités digitales.

Python propose des fonctions de calcul distribué via sa librairie **multiprocessing** [1]. Comme le cluster possède jusqu'à 48 coeurs (24 coeurs physiques, mais en tout 48 coeurs virtuels), il est possible de multiplier la vitesse d'exécution des différents programmes par 48.

Le processus de parallélisation se fait via des fonctions de type map-reduce. La fonction map va distribuer une tâche répétitive à tous les coeurs selon un vecteur input donné. La fonction reduce va rassembler et agréger les résultats obtenus. Dans certain cas, seule la fonction map

1. multiprocessing - https ://docs.python.org/2/library/multiprocessing.html

est nécessaire comme par exemple si les tâches sont indépendantes. C'est notamment le cas si la tâche à paralléliser est de calculer et insérer un certain nombre de données dans une table MySQL en interagissant directement avec la base de données (ce qui peut être fait par chacun des processus distribué).

Python possède toutefois certaines faiblesses au niveau de la rapidité d'exécution des algorithmes via son interpréteur, en particulier sur des boucles et la gestion des listes. Ainsi le calcul vectoriel et matriciel peut fortement ralentir l'exécution du code si celui ci ne tient pas compte de ces caractéristiques.

Plusieurs solutions existent pour compenser ces faiblesses. Il existe plusieurs librairies permettant le calcul vectoriel et matriciel comme **numpy**[2] et **scipy**[3]. Ces modules ont été particulièrement intéressants pour mener à bien les tâche de fitting de courbes. Cependant, la solution générale que nous avons choisie dans la plupart des cas (et malheureusement incompatible avec les modules numpy et scipy) est l'utilisation de **pypy**[4].

pypy permet l'exécution de code en langage python par un interpréteur écrit en langage C permettant une exécution du code plus rapide et compensant entre autres les faiblesses des boucles et listes du langage python. Cet outil ne nécessite aucune modification sur les algorithmes de base écrits en python, l'exécution du code en passant par pypy permet simplement de multiplier la vitesse d'exécution par un facteur dépendant du code (jusqu'a 40 fois plus rapide selon les algorithmes que nous avons utilisés).

Le seul facteur à considérer lors du passage à pypy est l'incompatibilité de certains modules existant en python. C'est notamment le cas pour numpy et scipy. Il existe aussi d'autres plateformes comme **anaconda**[5], plateforme open source de haute performance pour python. Anaconda est très utilisé en datascience et permet d'exécuter plusieurs versions de python dans des environnements détachés.

Divers modules de python se sont révélés extrêmement utiles notamment dans le traitement d'images nécessaire pour le calcul du chronocloud qui utilise la librairie **PIL**[6]. Cette librairie permet de manipuler aisément plusieurs images et de les combiner selon divers critères.

La librairie **pymysql**[7] est un module essentiel ayant permis d'exécuter toutes les interactions entre l'algorithme et la base de données MySQL y compris la construction et l'optimisation des index. Enfin, la librairie **BeautifulSoup**[8] a permit le parsing et le traitement des données textuelles depuis les fichiers au format XML.

2. numpy - http ://www.numpy.org
3. scipy - https ://www.scipy.org
4. pypy - https ://pypy.org
5. anaconda - https ://www.continuum.io/downloads
6. PIL - http ://effbot.org/imagingbook/pil-index.htm
7. pymysql - https ://github.com/PyMySQL/PyMySQL
8. BeautifulSoup - https ://www.crummy.com/software/BeautifulSoup/bs4/doc/

Analyse de n-grammes Partie IV

11 Analyse de 1-grammes

Dans ce chapitre, nous analysons le premier niveau des n-grammes, les 1-grammes. Il s'agit donc de l'étude diachronique des mots de l'ensemble du corpus. Ce niveau représente une unité de la langue parmi les plus élémentaires.

Dans le cadre d'études linguistiques de corpus, il est parfois utile d'utiliser un processus de transformation des données appelé lemmatisation. Ce processus permet d'identifier le lemme de chaque mot du corpus afin de les regrouper selon ce critère. Toutefois, nous avons choisi de ne pas effectuer cette transformation sur nos données pour deux raisons distinctes. La première est que cette transformation nous conduirait à perdre l'information de distinction des différentes formes des mots (genre, nombre, personne et mode) dans nos données alors que nous souhaitons la traiter de la même façon que les autres informations qui sont contenues dans ces données.

La seconde est que nous souhaitons développer des méthodes robustes et indépendantes des langues et corpus étudiés, considérant des suites de signes et leurs évolutions temporelles sans connaissance a priori de la langue étudiée. En effet, procéder à une lemmatisation demande une connaissance préétablie de la langue dans laquelle ont été écrits les textes composant le corpus et n'a donc pas vocation à être utilisée dans le contexte de cette étude.

Ce chapitre est subdivisé selon l'utilisation des outils et concepts développés dans les chapitres précédents. Il débute par les concepts du niveau Macro et tend progressivement vers ceux du niveau Micro, permettant d'observer le corpus d'abord de façon globale et distante pour ensuite l'examiner de plus en plus près et selon différents angles. Le chapitre commence donc par l'étude de l'évolution des distances et de l'entropie. Nous utilisons ensuite la visualisation chronocloud pour identifier les mots fréquents du corpus et appréhender son contenu global à travers l'évolution temporelle des mots de résilience $50 \leqslant R < 150$. Nous étudions ensuite l'évolution de mots particuliers au travers de l'outil de visualisation des profils fréquentiels, le visualisateur de n-grammes. Enfin le chapitre se termine par une discussion des résultats obtenus ainsi que des avantages et désavantages des méthodes utilisées.

11.1 Analyse diachronique des distances

Dans la partie précédente, nous avons constaté dans les corpus de JDG et GDL des perturbations non linguistiques potentielles comme l'augmentation de la fréquence des mots représentant des nombres en raison de l'intégration de section dédiées aux horaires des trains et du cinéma ainsi que des données financières et boursières. En outre, les corpus contiennent des erreurs d'OCR dont la proportion varie avec les années. Ces corpus peuvent facilement être divisés en sous-corpus correspondant à l'année de publication. Cependant, le nombre de pages et de mots fluctuent considérablement en fonction des années, allant de 280 000 mots par année au début du 19e siècle à environ 18 millions dans les dernières années du XXe siècle. La Figure 11.1 présente l'évolution de la taille des corpus en nombre de mots par année pour JDG et GDL.

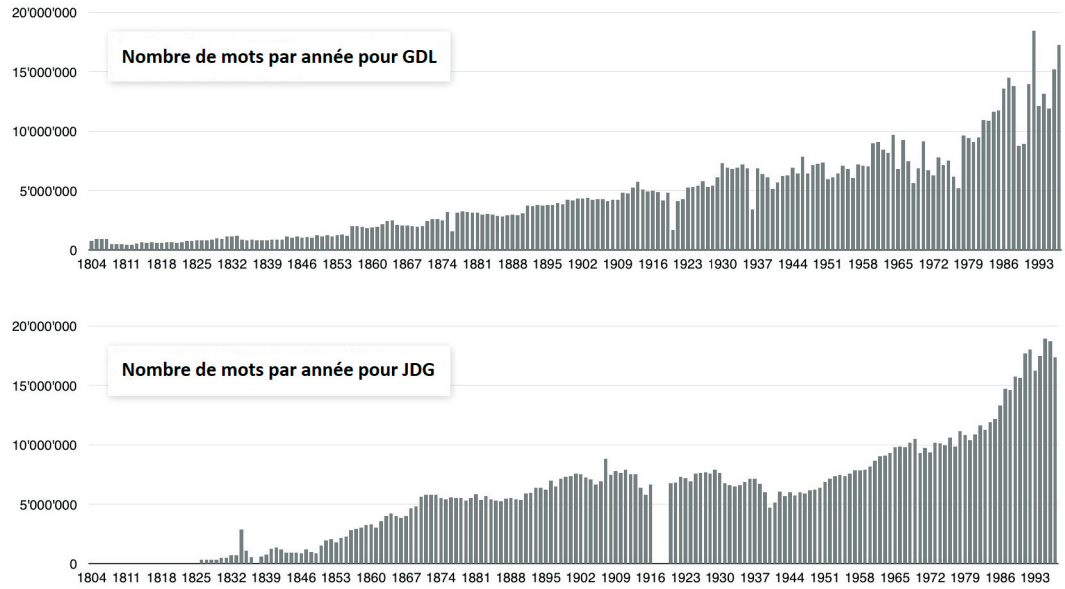


FIGURE 11.1 – Taille du corpus par années pour GDL (en haut) et JDG (en bas)

De plus, certaines périodes comme celle de 1900 à 1915 pour JDG et celle de 1965 à 1998 pour les deux journaux présentent des niveaux de bruit plus élevés que les autres. S'il est habituel d'appliquer un filtre de fréquence pour gérer ce type de problème, nous choisissons d'utiliser également les notions de noyau et distance nucléaire en plus de méthodes classiques afin de tester la robustesse de ces méthodes et tenter de mesurer les changements linguistiques tout en évitant d'éventuelles mauvaises interprétations dues aux fluctuations de bruit et aux variations de taille du corpus. Compte tenu du manque de données pour les années 1837, 1917, 1918 et 1919 de JDG, nous les avons retirées des graphiques et analyses. En outre, certaines années ont également été écartées en raison de la qualité médiocre des journaux scannés (1834, 1835, 1859 et 1860 pour JDG et 1808 pour GDL).

Distance de Jaccard

Nous comparons d'abord les sous-corpus annuels de JDG et GDL à l'aide de la distance de Jaccard qui mesure, pour deux corpus C_1 et C_2, la différence entre leurs lexiques $L_1 = L(C_1)$ et $L_2 = L(C_2)$. Nous appliquons un filtre fréquentiel afin de ne conserver dans les lexiques que les mots dont la fréquence relative est supérieure à $1/100000$. Il est possible de choisir un filtre basé sur le nombre d'occurrences, mais la fréquence relative présente l'avantage de tenir compte des différences de taille des corpus. Toutefois, les données filtrées peuvent toujours présenter des erreurs d'OCR et du bruit. En effet, un seuil de fréquence trop élevé entraine la perte de données intéressantes tandis qu'à l'inverse un seuil trop faible laisse trop de bruit et d'erreurs contaminer l'analyse du corpus. Le calcul de la distance de Jaccard $d(L_1, L_2) = 1 - \frac{|L_1 \cap L_2|}{|L_1 \cup L_2|}$ comparant deux à deux tous les sous-corpus annuels donne une matrice symétrique $M \times M$ où M est le nombre d'années distinctes du journal étudié. Cette matrice contient toutes les distances entre chaque paires d'années i et j (donnant la comparaison entre $L(C_i)$ et $L(C_j)$) normalisées dans l'intervalle $[0, 1]$.

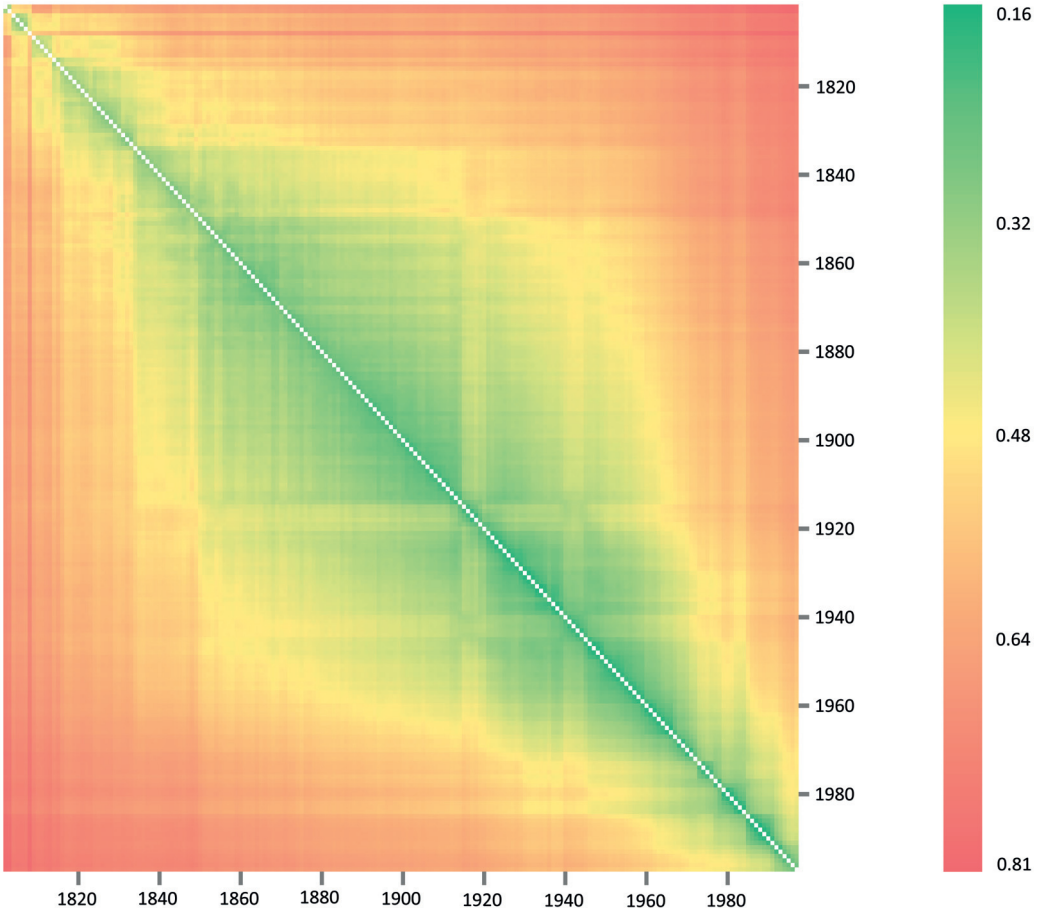


FIGURE 11.2 – Heatmap de la matrice de distances de Jaccard pour le corpus de GDL

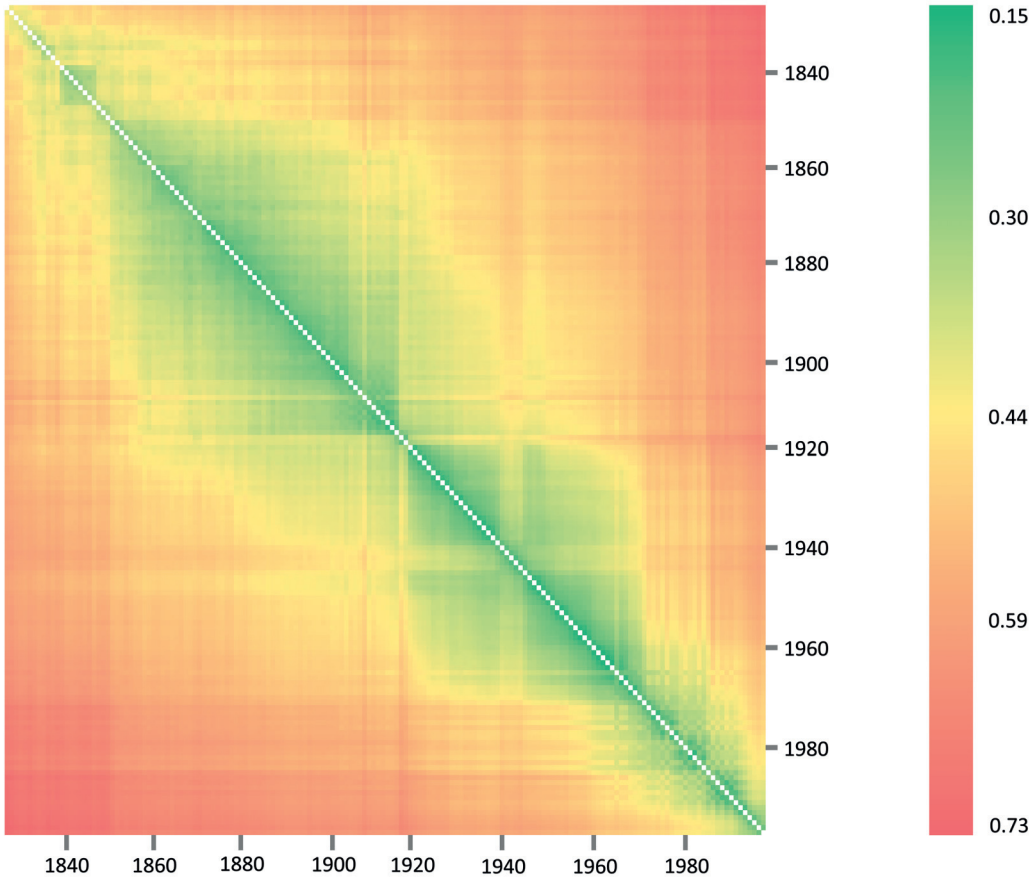


FIGURE 11.3 – Heatmap de la matrice de distances de Jaccard pour le corpus de JDG

les Figures 11.2 et 11.3 présentent une visualisation de type heatmaps de la matrice de distances de Jaccard pour les corpus de GDL et JDG. Comme attendu, nous observons que les valeurs de la matrice sont fortement corrélées avec la différence entre les années comparées. De plus, les lignes de niveau de la carte suggèrent que l'évolution des corpus n'est pas linéaire, mais plutôt structurée par période de temps. En effet, dans le cas d'une évolution linéaire, les lignes de niveau seraient parallèles à la diagonale. Or nous observons plutôt quatre périodes clairement séparées pour les deux corpus, le début du corpus à 1849, 1850 à 1915, 1916 à 1971 et 1972 à la fin du corpus. La séparation entre la deuxième et troisième période est moins claire dans le corpus de GDL. Ainsi, chaque période semble possèder un régime propre de sorte que les sous-corpus composant une période donnée sont généralement plus proches.

Nous avons représenté dans les Figures 11.4 et 11.5, une partie de cette matrice dans un graphique montrant les valeurs de la distance en fonction de la différences de temps entre les sous-corpus. Dans cette représentation, nous observons que les distances sont globalement proportionnelles au nombre d'années séparant les deux sous-corpus. Cette observation suggère que le changement linguistique existe et peut être quantifiée par la distance de Jaccard.

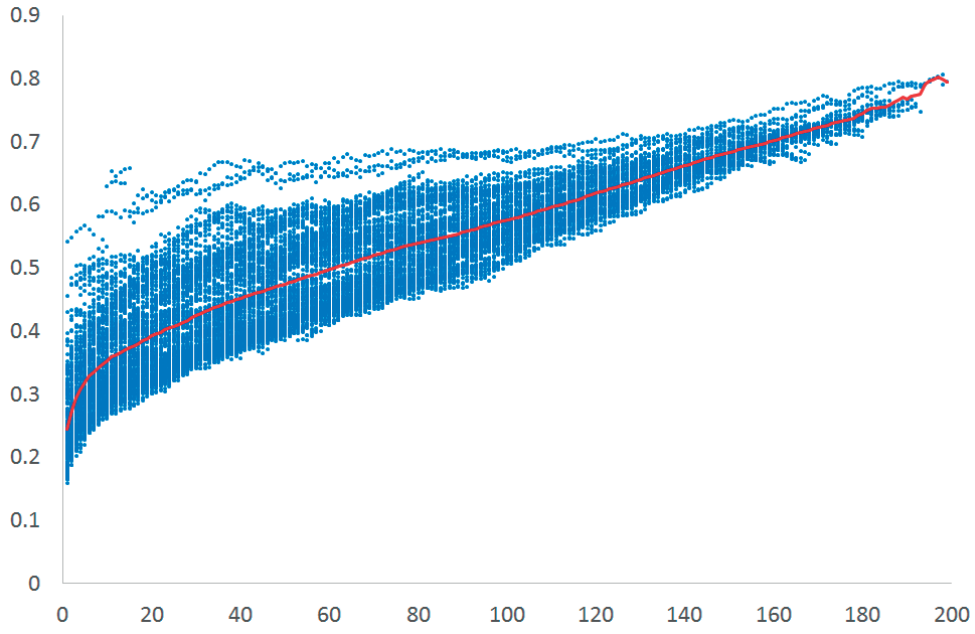


FIGURE 11.4 – Distances de Jaccard (bleu) et moyenne de ces distances (rouge) en fonction du nombre d'années de différence entre les sous-corpus de GDL

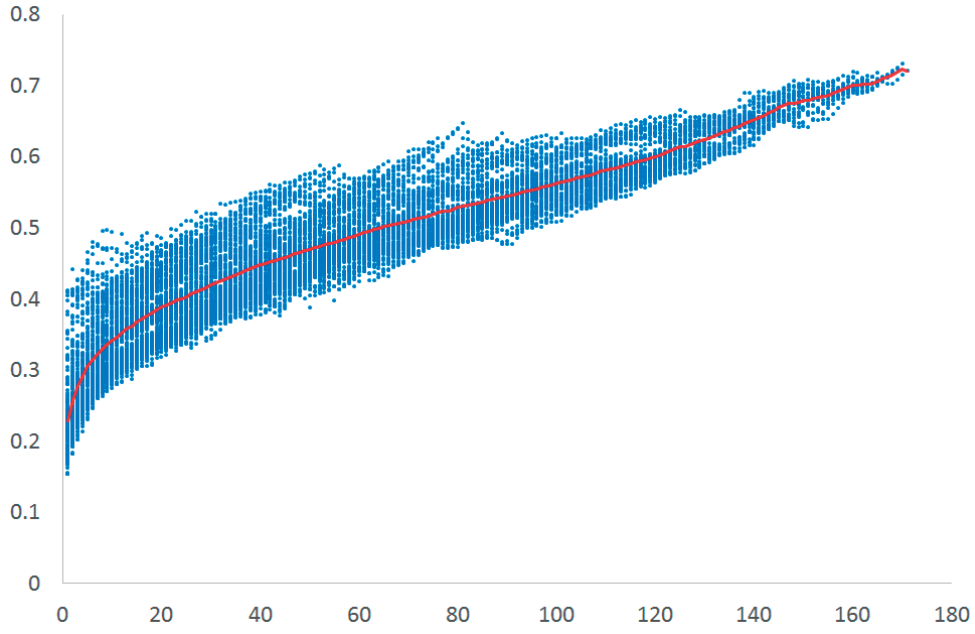


FIGURE 11.5 – Distances de Jaccard (bleu) et moyenne de ces distances (rouge) en fonction du nombre d'années de différence entre les sous-corpus de JDG

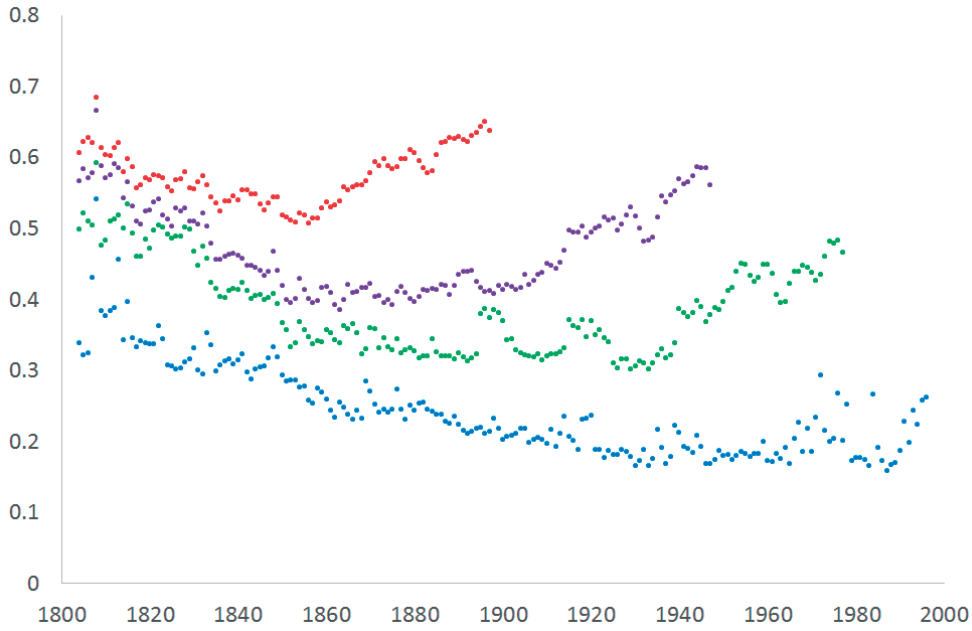


FIGURE 11.6 – Distances de Jaccard entre les années y_i et y_{i+n} avec $n = 1$ (bleu), $n = 20$ (vert), $n = 50$ (violet) and $n = 100$ (rouge) pour le corpus de GDL

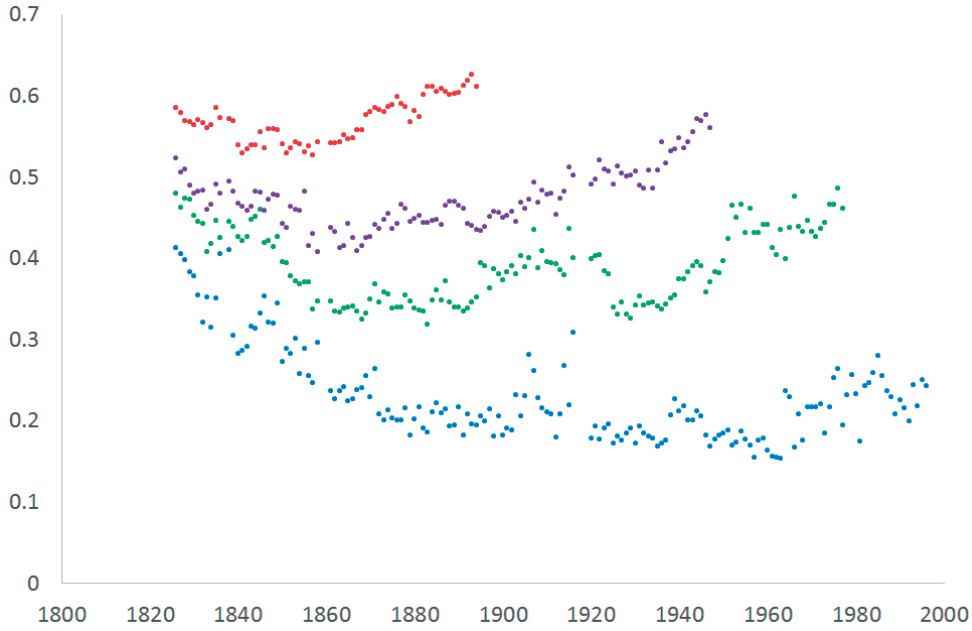


FIGURE 11.7 – Distances de Jaccard entre les années y_i et y_{i+n} avec $n = 1$ (bleu), $n = 20$ (vert), $n = 50$ (violet) and $n = 100$ (rouge) pour le corpus de JDG

La corrélation observée dans la Figure 11.1 entre la taille du corpus et le temps pourrait affecter la valeur de la distance au même titre que la quantité de temps séparant les sous-corpus. Toutefois, si nous restreignons les calculs de la matrice de distances de Jaccard aux données des années les plus stables en termes de taille et recalculons la visualisation représentée dans les Figures 11.4 et 11.5, nous observons globalement la même évolution de la distance de Jaccard. Nous remarquons dans ces figures que le comportement de la moyenne des distances (courbe rouge) est plus sensible aux premières années de séparation pour les deux journaux.

Afin de mesurer l'évolution de ces changements linguistiques et de déterminer s'ils accélèrent, ralentissent ou restent stables, nous représentons une visualisation partielle de la matrice de distance en représentant uniquement les distances entre les années y_i et y_{i+n} avec n égal à 1, 20, 50 et 100, ces visualisations sont présentées dans les Figures 11.6 et 11.7 pour les corpus de GDL et JDG. Nous observons que la distance entre une année et la suivante diminue lentement avant de se stabiliser à partir de l'année 1920. Cela suggère que la langue est plus stable à partir de cette année. Nous observons toutefois une instabilité marquée après 1965, correspondant à la période dont le niveau de bruit est plus élevé.

Pour la distance $d(y_i, y_{i+n})$, nous observons que l'évolution de la distance dont la différence est de 20 années diminue lentement avant d'augmenter de façon significative dans les dernières années. Il est possible que cette augmentation soit due à la "contamination" de la matrice de distance par les données correspondant aux années perturbées par diverses données non linguistiques (nombres, bourse, horaires, cinéma). Les mêmes graphiques sans cette période de bruit ne montrent pas la moindre augmentation de la distance et il est donc peu prudent d'interpréter cela comme une accélération de l'évolution linguistique du corpus.

La matrice de distance de Jaccard indique un effet global qui pourrait être causé par un changement linguistique, comprenant le processus d'apparition de nouveaux mots et la disparition de mots anciens. Cependant, la distance de Jaccard est connue pour être affectée par les grandes différences de taille de corpus (Muller, 1980), et d'autres définitions de distances ont été conçues pour corriger cette propriété indésirable. Une distance de Jaccard améliorée est présentée dans une étude des similarités des textes (Brunet, 2003) dans le but d'éliminer ou réduire la sensibilité de la distance à la différence de taille du corpus. Nous avons calculé la distance de Jaccard améliorée sur les corpus de GDL et JDG et il semble que cette distance ait le même comportement que la distance de Jaccard classique, mais normalisée différemment.

De plus, les erreurs d'OCR et le bruit affectent la distance de Jaccard en raison de sa nature binaire et l'absence de considération de la fréquence des mots. Les filtres de fréquence peuvent être utilisés pour diminuer l'influence du bruit, mais le seuil appliqué est assez arbitraire et difficile à justifier. En effet, nous avons opté pour un seuil de fréquence relative fixé à 1/100000 afin de balancer l'effet de la perte de données potentiellement intéressantes avec celui d'un taux trop important de bruit et d'erreurs d'OCR. Une autre option possible et moins arbitraire est un simple seuil d'occurrence fixé à la valeur 1. Toutefois, un tel seuil induirait des inégalités en fonction de la taille des données et donc en fonction du temps.

Nous présentons dans la Figure 11.8 un tableau récapitulatif des six Figures 11.2 à 11.7 afin de permettre la comparaison des corpus de GDL et JDG.

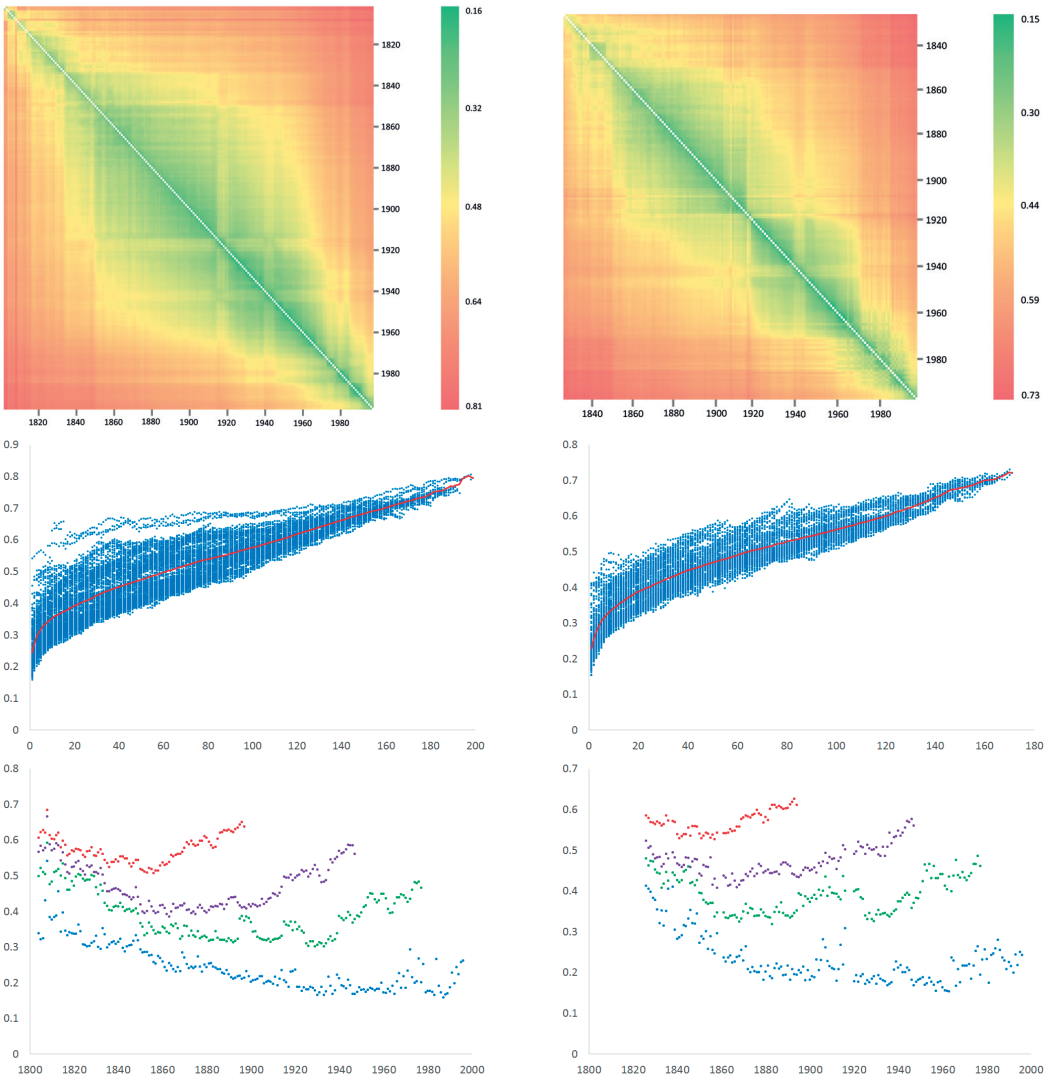


FIGURE 11.8 – **(1)** : GDL ; **(2)** : JDG ; **Haut** : Heatmap de la matrice des distances de Jaccard ; **Milieu** : Distances de Jaccard (bleu) et moyenne de ces distances (rouge) en fonction du nombre d'années de différence entre les sous-corpus ; **Bas** : Distances de Jaccard entre les années y_i et y_{i+n} avec $n = 1$ (bleu), $n = 20$ (vert), $n = 50$ (violet) and $n = 100$ (rouge)

La répartition inégale de la taille des sous-corpus provoque donc des difficultés méthodologiques pour interpréter les distances classiques comme celle de Jaccard. Les fluctuations de la taille et le bruit provoquent une augmentation indirecte de la mesure du changement linguistique par la formule de Jaccard. Dans de telles conditions, il est difficile de distinguer les effets des variations de taille du corpus de ceux de l'apparition réelle des nouveaux mots dans le lexique de la langue ainsi que de la disparition des anciens mots.

Ces difficultés d'interprétation motivent l'exploration d'une autre approche par le biais des notions de noyau résilient et d'ensembles résilients.

Nous considérons les noyaux résilients, $K_{1804,1997,GDL}$ contenant 5 242 mots uniques qui ont été utilisés durant environ 200 ans dans GDL et $K_{1826,1997,JDG}$ contenant 7 486 mots uniques couvrant une période d'environ 170 ans dans JDG. La Figure 11.9 donne la composition des noyaux résilients de GDL et JDG en termes de partie du discours, obtenue avec TreeTagger (Schmid, 1995) avec une visualisation de type piechart et barchart.

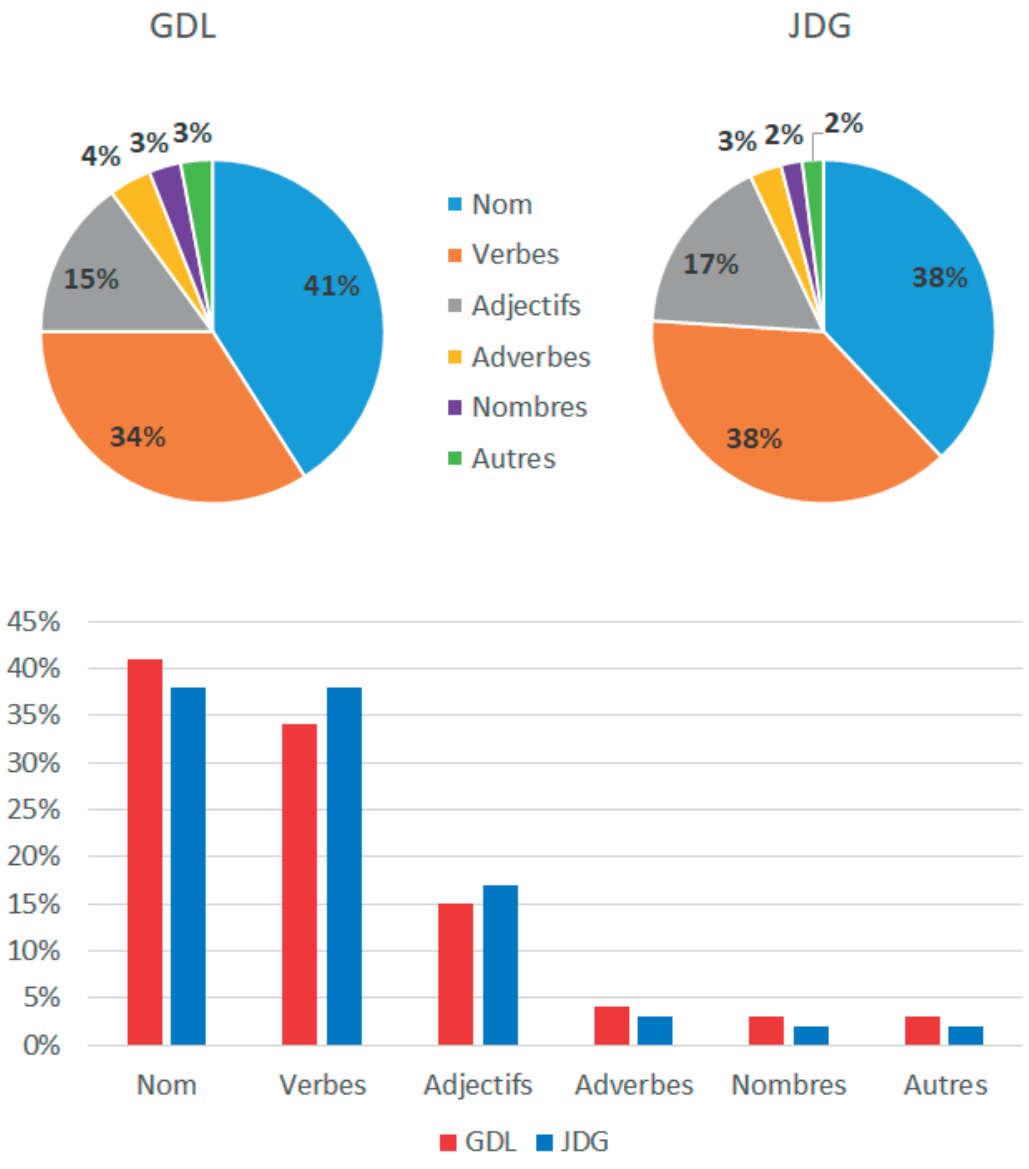


FIGURE 11.9 – Distribution des parties du discours dans le noyau résilient. **(1)** : GDL ; **(2)** : JDG ; **Haut** : Visualisation de type piechart ; **Bas** : Visualisation de type barchart

Nous observons une répartition similaire des parties du discours dans les noyaux résilients de GDL et JDG. Il est toutefois intéressant de constater que le noyau de GDL contient plus de noms et le noyau de JDG plus de verbes.

Ces deux noyaux possèdent 4 464 mots en commun soit 85% du noyau de GDL et 60% du noyau de JDG. Cette différence n'est pas anormale, car le noyau de GDL couvre une plus large période que celui de JDG, l'intersection de ces deux ensembles est donc principalement conditionnée par le corpus GDL.

Soit l'ensemble résilient $R_{x,C}$, qui contient tous les mots qui se maintiennent dans le corpus C pendant au moins x années. Les ensembles résilients peuvent être organisés sous forme d'ensembles concentriques (cf. chronocloud) avec la propriété $R_{1,C} \subset R_{2,C} \subset ... \subset R_{i,C} \subset R_{i+1,C}$.

La proportion relative de chaque sous-ensemble donne des informations sur la stabilité et la dynamique du changement de la langue. La Figure 11.10 montre l'évolution du nombre d'éléments dans les ensembles résilients pour les deux journaux.

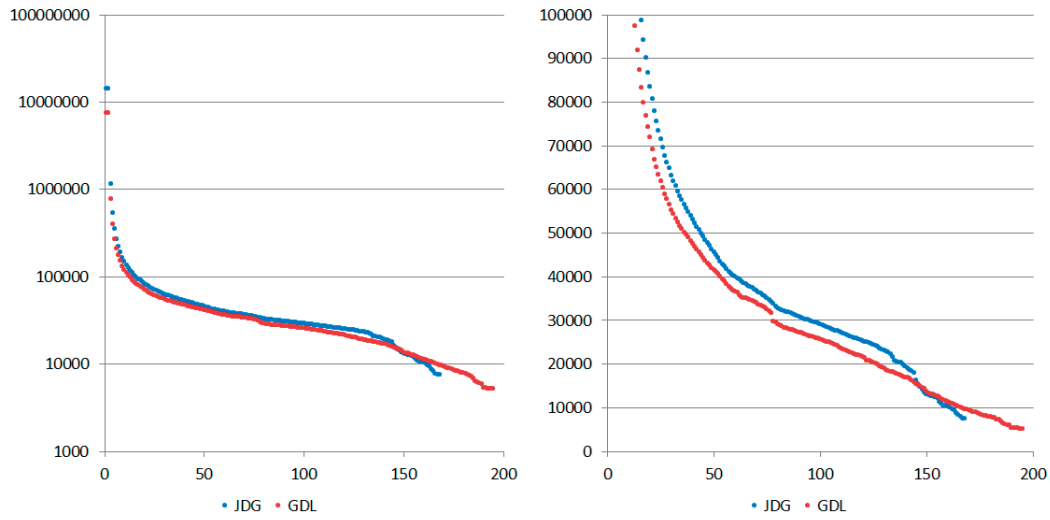


FIGURE 11.10 – Taille des ensembles résilients R_d en fonction de la résilience d pour GDL (rouge) et JDG (bleu). **(1)** : échelle logarithmique ; **(2)** : échelle linéaire

Cette représentation de R_d montre une tendance générale concernant la taille des ensembles résilients pour les corpus de JDG et GDL. Ces deux courbes sont similaires bien qu'un croisement ait lieu vers la résilience de 150 années.

Dans le but de comparer les deux noyaux sur une base plus équitable en terme d'années de couverture, nous calculons un noyau de GDL réduit aux années couvertes par le corpus de JDG. Nous avons alors un noyau GDL contenant 8 975 mots à la place de 4 464. Le corpus de GDL a donc un vocabulaire résilient plus large que celui de JDG. Le nombre de mots communs aux deux noyaux devient alors 6 582, soit 73% du noyau de GDL et 88% du noyau de JDG.

Afin d'analyser de plus près la composition des ensembles résilients, nous calculons le nombre de mots en fonction de leur résilience. Ce nombre de mots dans le corpus C et pour une résilience x vaut $\|R_{x,C}\| - \|R_{x-1,C}\|$. Nous présentons le nombre de mots en fonction de leur résilience en échelle logarithmique dans la Figure 11.11.

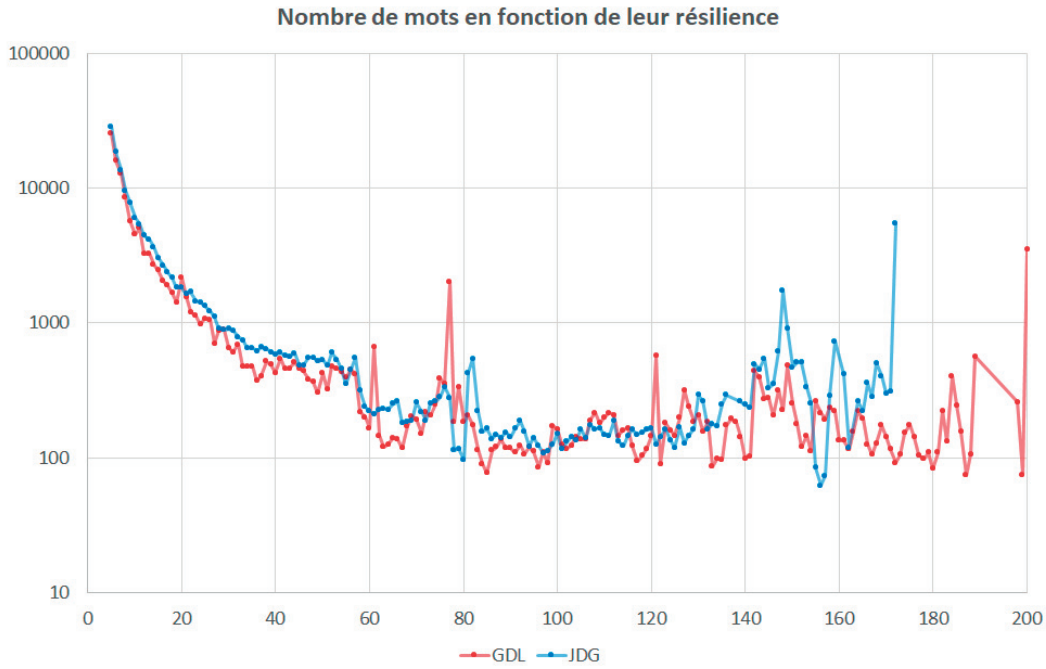


FIGURE 11.11 – Nombre de mots en fonction de leur résilience en échelle logarithmique

Nous observons certains sursauts notamment pour des résiliences particulières comme 61, 77, 121, 184 pour GDL et 78, 82, 148, 156, 159 pour JDG. Nous observons aussi les derniers sursauts qui correspondent au noyaux résilients des journaux.

Chaque résilience correspond à la durée maximale pour laquelle un mot est présent chaque année consécutive. On peut donc affiner cette analyse en considérant l'année de début de cette période maximale et l'année de fin de cette même période pour chaque mot. Ces notions correspondent respectivement à une définition raisonnable de l'année d'apparition du mot et son année de disparition.

Définition 15. *L'année d'apparition d'un mot dans un corpus donné est l'année la plus ancienne de sa période de résilience maximale dans ce corpus*

Définition 16. *L'année de disparition d'un mot dans un corpus donné est l'année la moins ancienne de sa période de résilience maximale dans ce corpus*

Sur cette base, nous pouvons observer le nombre de disparitions et d'apparitions de mots par année présentés dans la Figure 11.12.

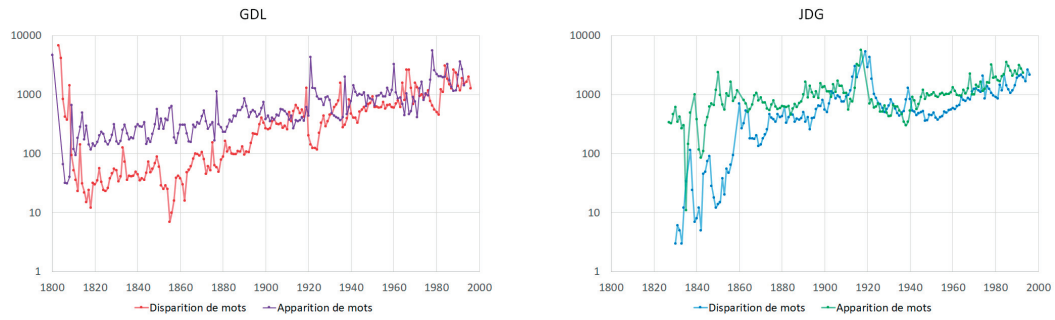


FIGURE 11.12 – Disparition (rouge / bleu) et apparition (violet / vert) de mots par année pour GDL (gauche) et JDG (droite)

Nous observons une apparition de mots généralement plus élevée que pour la disparition. Les écarts tendent à se réduire avec les années bien que certaines périodes montrent des écarts élevés avec un nombre de disparitions diminuant et d'apparitions augmentant. La période de la première guerre mondiale n'a pas le même comportement au sein des deux journaux.

Le manque de données de JDG dans cette période n'est pas dommageable dans le sens où ces années sont simplement ignorées, d'où un court décalage. Toutefois, le corpus de GDL montre un écart se creusant entre les courbes d'apparition et disparition de mots dans cette période. Le corpus de JDG montre un écart également important dans la période de 1940 à 1960.

Ces mêmes courbes sont présentées dans la Figure 11.13 dans une configuration qui facilite la comparaison des corpus.

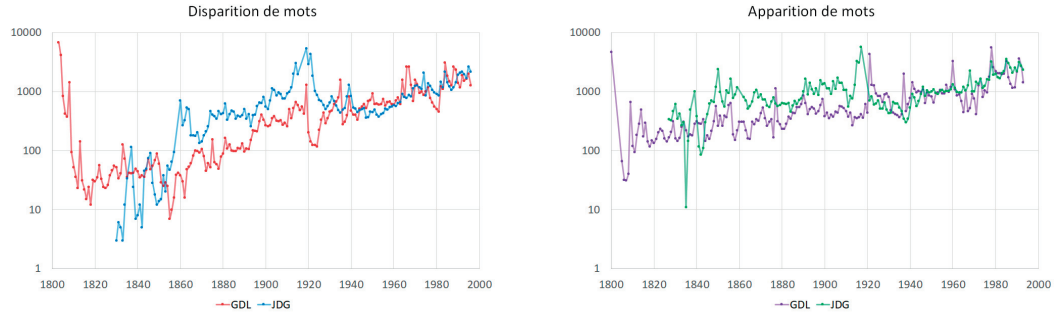


FIGURE 11.13 – Disparition (gauche) et apparition (droite) de mots par année pour GDL (rouge / violet) et JDG (bleu / vert)

Nous observons que le nombre de disparitions et apparitions de mots est supérieur dans JDG durant la période allant de 1860 à 1920-1925. Durant cette période les courbes montrent des comportements parfois opposés, toutefois elles ont une tendance à se rejoindre avec les années. A partir d'environ 1930, les deux journaux ont des valeurs similaires. Nous présentons les informations complètes de la matrice apparitions/disparitions via une visualisation de type heatmap présentée dans la Figure 11.14.

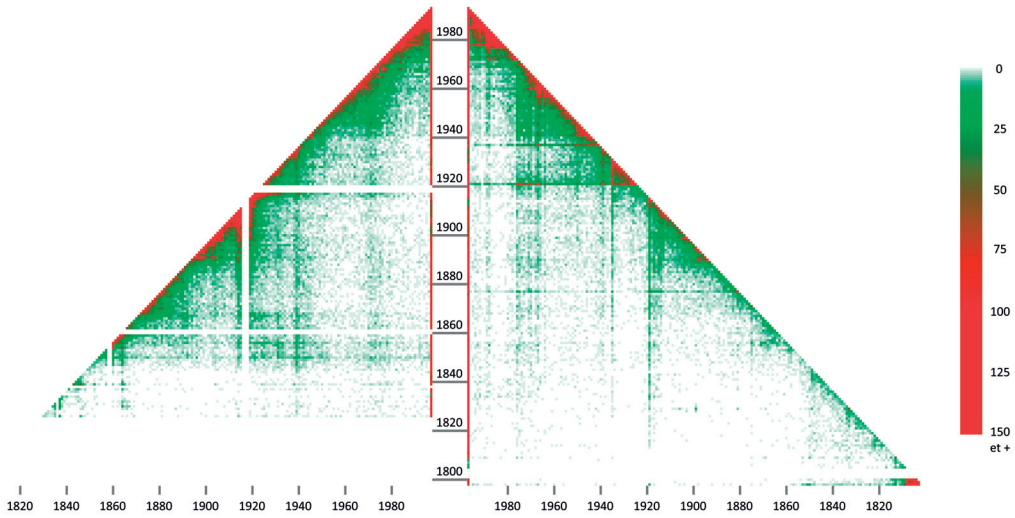


FIGURE 11.14 – Années d'apparition (axe vertical) et de disparition (axe horizontal) de mots pour JDG (gauche) et GDL (droite), la couleur donnant le nombre de mots

Nous observons plusieurs effets sur cette visualisation :

- Les raisons des sursauts constatés dans la Figure 11.11 sont essentiellement des effets de bord dus au nombre élevé de mots déjà présents avant le début du corpus ou présents à la fin de celui-ci. Une année rattachée à des données perturbées aura alors pour effet de sur-représenter une certaine valeur de la résilience.
- Les deux lignes verticales rouges montrent les mots qui survivent jusqu'à la fin du corpus. Le noyau résilient de chaque corpus correspond au point le plus bas de cette ligne. Cela suggère l'apparition de nombreux mots dans le corpus qui sont potentiellement résilients, mais non inclus dans le noyau résilient par définition.
- Les effets des deux guerres mondiales sur GDL se manifestent par les lignes de couleur vertes signifiant une augmentation de l'activité d'apparition et de disparition des mots. Sur JDG, les années 1917, 1918 et 1919 sont manquantes, mais une agglomération de mots instables est constatée autour de ces dates.
- Le tout début du corpus de GDL exhibe des mots instables potentiellement dus aux nombreuses erreurs d'OCR rattachées à ces années.
- En lien manifeste avec l'augmentation de la taille du corpus, nous constatons que les années les plus récentes ont une activité d'apparition et de disparition de mots plus élevée.

Le noyau résilient correspond donc à une part infime de l'ensemble des mots. Toutefois, ces mots ont une présence significative au sein de la langue et leur étude diachronique peut révéler les changements linguistiques du corpus. Pour cela, nous utilisons la notion de distance nucléaire définie sur le noyau résilient afin d'étudier des effets linguistiques indépendamment des problèmes liés à la nature du corpus (évolution de layout, erreurs d'OCR, perturbations numériques ou autres instabilités non linguistiques).

Distance nucléaire

Une fois les noyaux déterminés, nous appliquons le calcul de la distance nucléaire au noyau résilient de chaque corpus. Cette distance d_{12}^K entre deux sous-corpus correspondant aux périodes 1 et 2 est appliquée sur le noyau résilient K via l'indice $I_t(w)$ correspondant à la position de l'élément w dans K ordonné selon les fréquences de la période t. Nous rappelons sa formule :

$$d_{12}^K = \frac{1}{M} \sum_{w \subset K} |I_1(w) - I_2(w)|$$

Sa normalisation M se calcule sur deux ensembles de même taille $N = \|K\|$, mais dont les éléments ont un ordre inverse. Nous rappelons également sa formule :

$$M = \sum_{i=1}^{N} |i - (N+1-i)| = \sum_{i=1}^{N} |2i - N + 1|$$

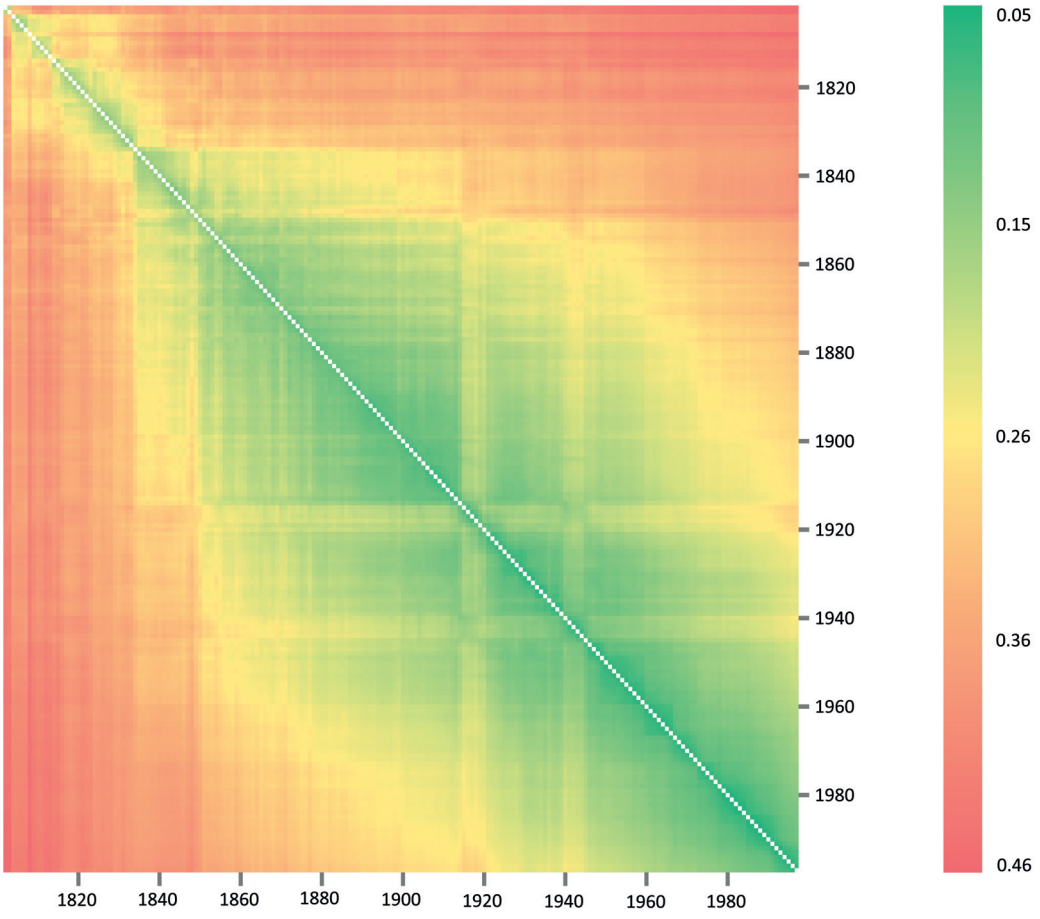


FIGURE 11.15 – Heatmap de la matrice des distances nucléaires pour le corpus de GDL

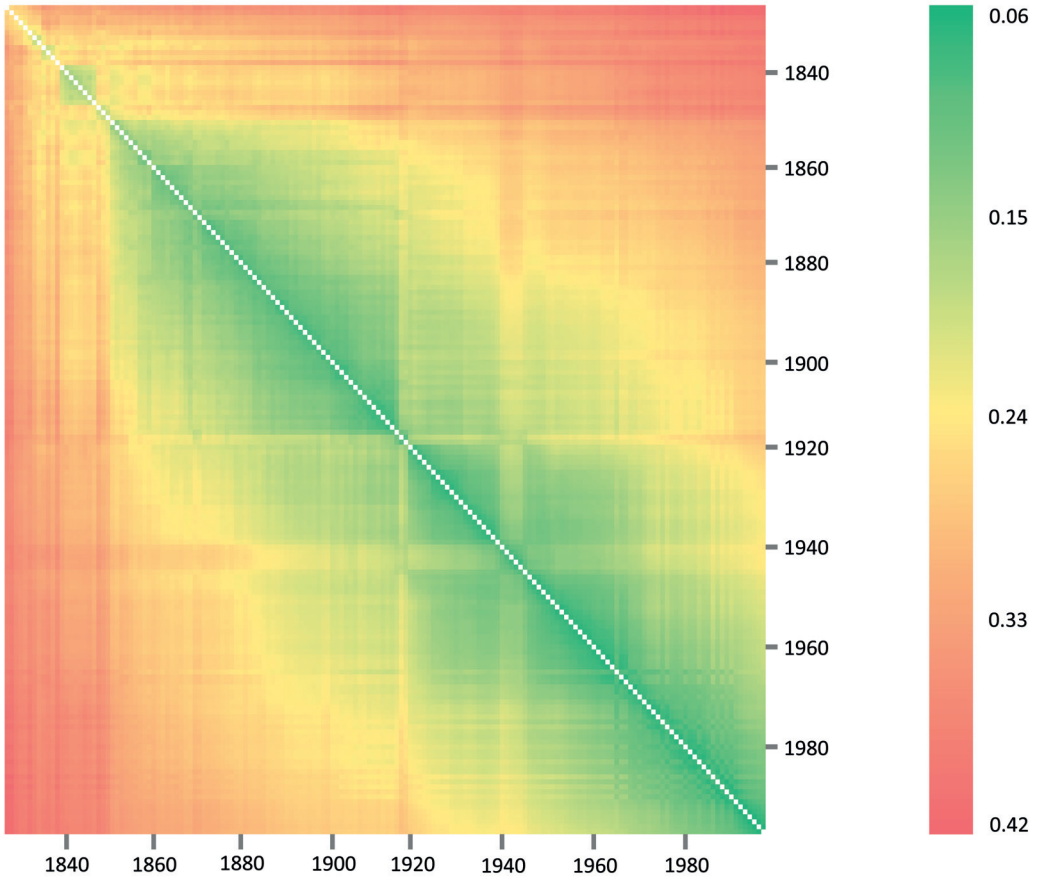


FIGURE 11.16 – Heatmap de la matrice des distances nucléaires pour le corpus de JDG

Nous avons ensuite généré des visualisations de la matrice de distances nucléaires similaires à celles utilisées avec la matrice de distances de Jaccard afin de pouvoir les comparer.

La visualisation de type heatmap est présentée dans les Figures 11.15 et 11.16. Les distances entre les années y_i et y_{i+n} en fonction de n sont présentées dans les Figures 11.17 et 11.18. Les distances entre les années y_i et y_{i+n} avec $n = 1$, $n = 20$, $n = 50$ et $n = 100$ en fonction de y_i sont présentées dans les Figures 11.19 et 11.20.

Ces figures sont présentées selon différentes échelles afin de visualiser principalement les différences de comportement plutôt que la normalisation inhérente qui diffère pour les deux distances. La distance de Jaccard s'étend de 0.16 à 0.81 pour GDL et de 0.15 à 0.73 pour JDG. La distance nucléaire s'étend de 0.05 à 0.46 pour GDL et de 0.05 à 0.42 pour JDG.

Le résumé de ces visualisations est présenté dans la Figure 11.21 ainsi qu'un résumé global comparant toute les figures pour les deux distances et les deux journaux dans la Figure 11.22.

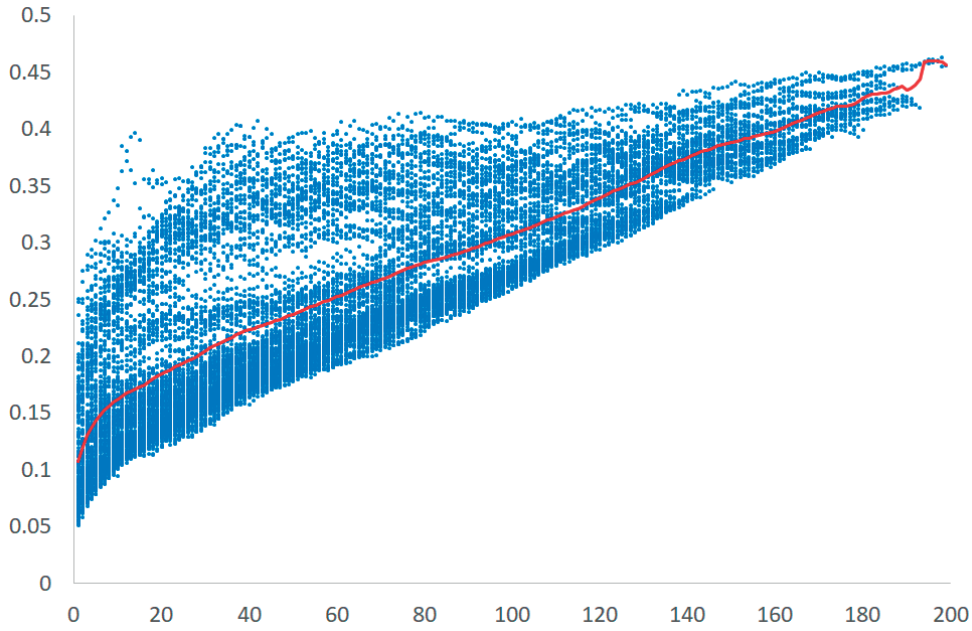


FIGURE 11.17 – Distances nucléaires (bleu) et moyenne de ces distances (rouge) en fonction du nombre d'années de différence entre les sous-corpus du corpus de GDL

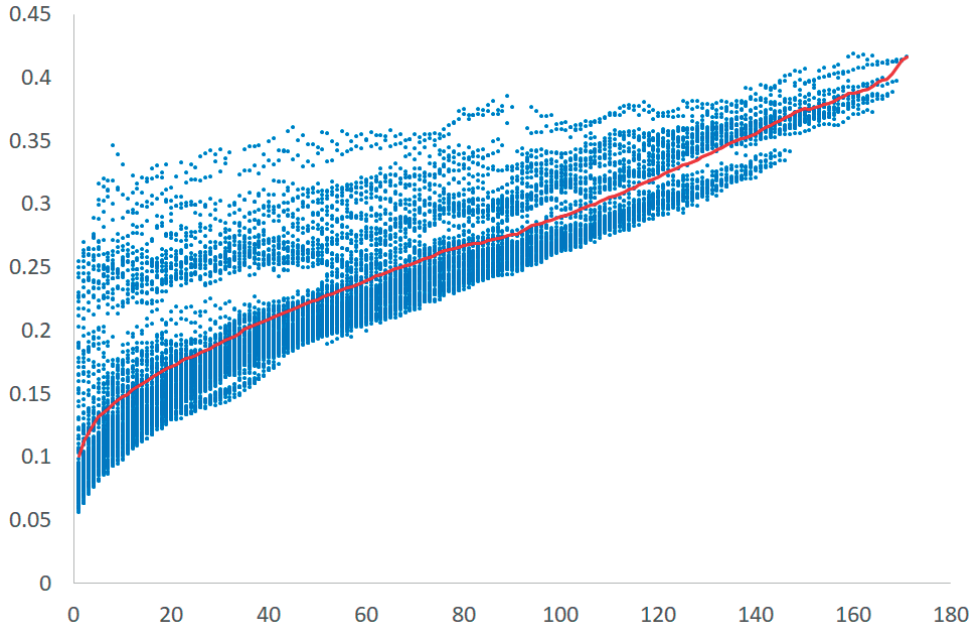


FIGURE 11.18 – Distances nucléaires (bleu) et moyenne de ces distances (rouge) en fonction du nombre d'années de différence entre les sous-corpus du corpus de JDG

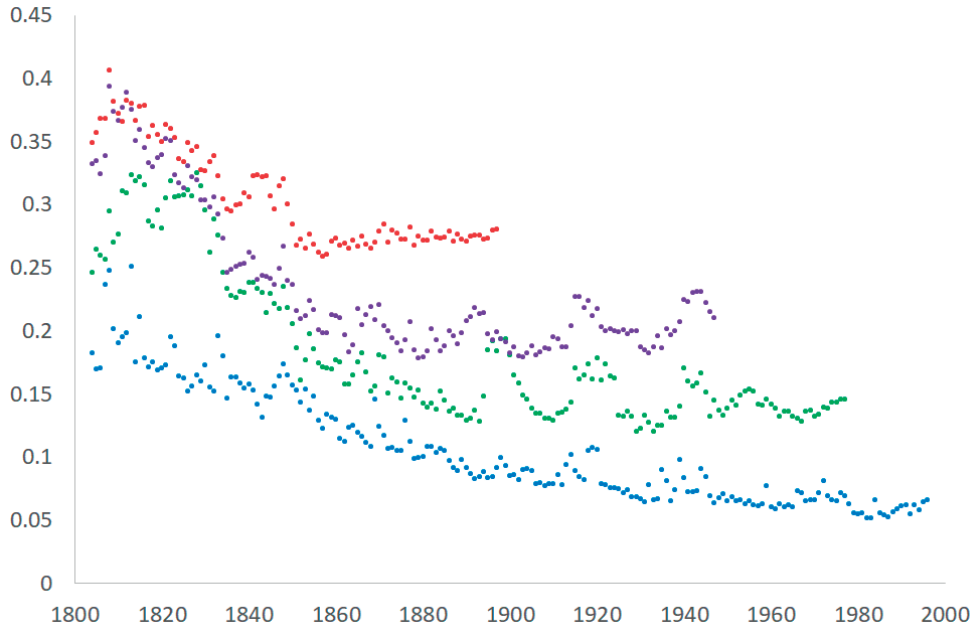


FIGURE 11.19 – Distances nucléaires entre les années y_i et y_{i+n} avec $n = 1$ (bleu), $n = 20$ (vert), $n = 50$ (violet) and $n = 100$ (rouge) pour le corpus de GDL

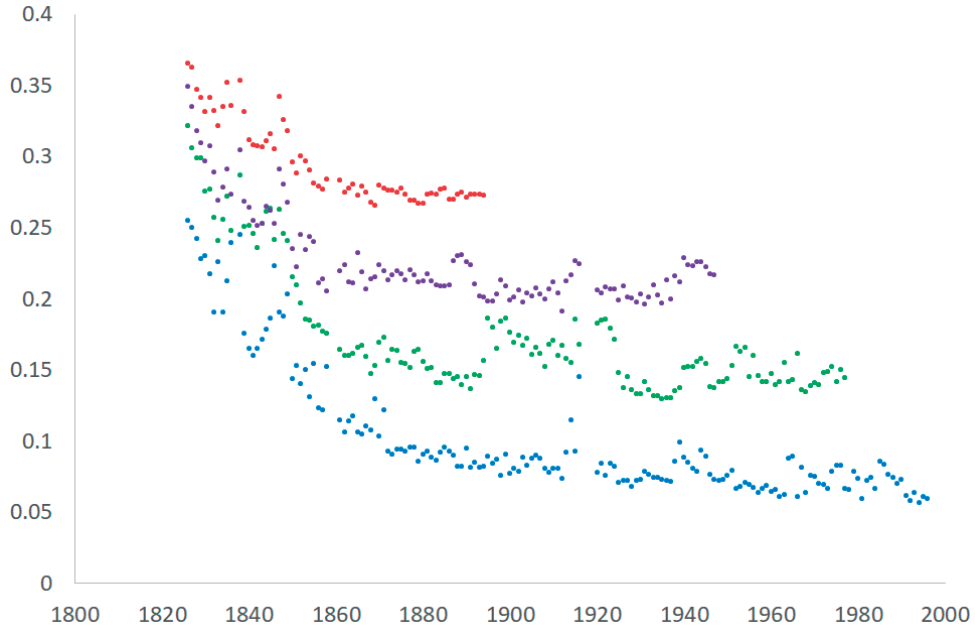


FIGURE 11.20 – Distances nucléaires entre les années y_i et y_{i+n} avec $n = 1$ (bleu), $n = 20$ (vert), $n = 50$ (violet) and $n = 100$ (rouge) pour le corpus de JDG

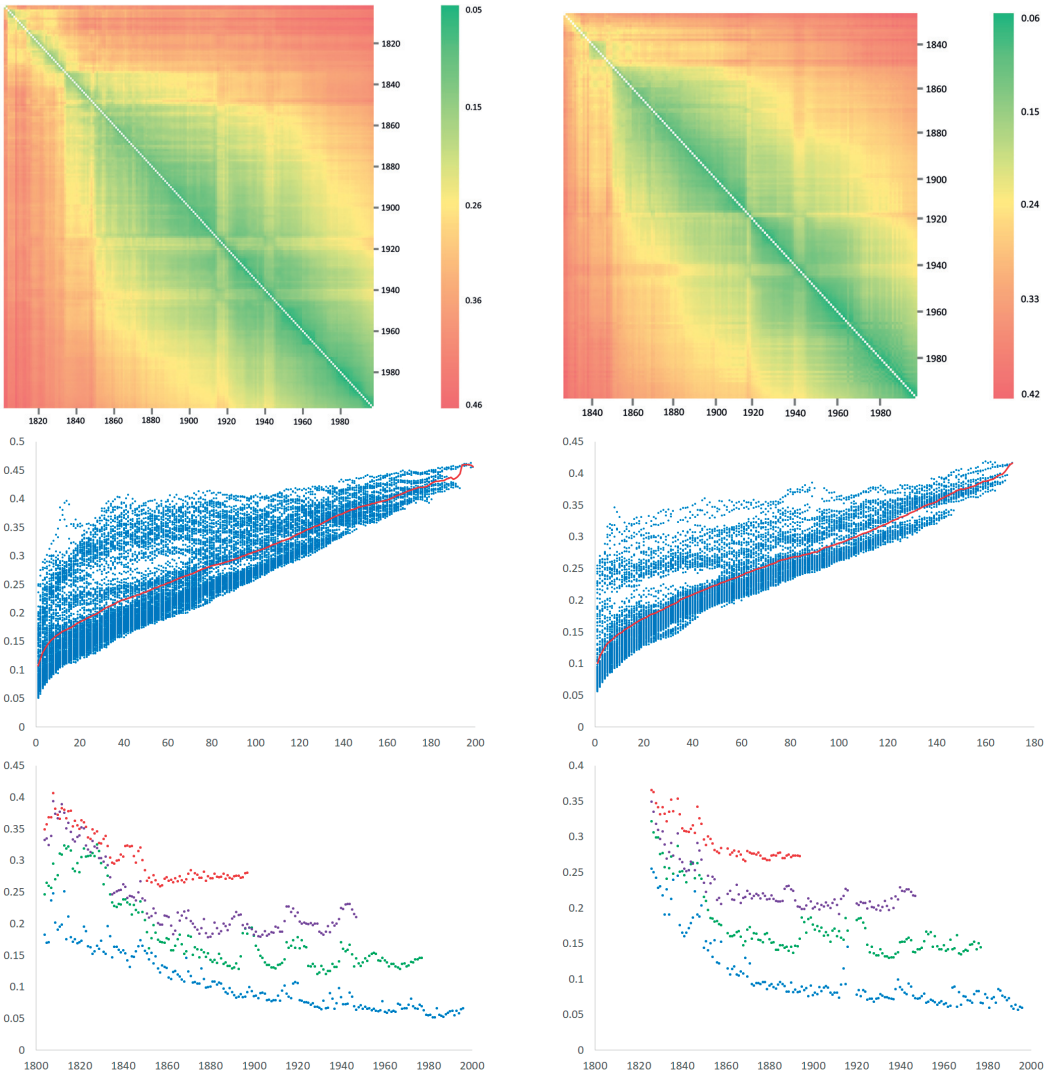


FIGURE 11.21 – **(1)** : GDL ; **(2)** : JDG ; **Haut** : Heatmap de la matrice des distances nucléaires ; **Milieu** : Distances nucléaires (bleu) et moyenne de ces distances (rouge) en fonction du nombre d'années de différence entre les sous-corpus ; **Bas** : Distances nucléaires entre les années y_i et y_{i+n} avec $n = 1$ (bleu), $n = 20$ (vert), $n = 50$ (violet) and $n = 100$ (rouge)

La distance de Jaccard et la distance nucléaire sont basées sur des éléments totalement différents par définition, la première est basée sur la présence et l'absence de mots dans les lexiques comparés et la seconde sur la comparaison de l'ordre en fréquence des mots qui sont justement commun à tous les lexiques.

Les deux distances augmentent avec la différence de temps entre les sous-corpus, soutenant l'hypothèse selon laquelle le changement linguistique du corpus existe et est quantifiable. La valeur de la distance de Jaccard va de $0,25$ à $0,8$ environ. La valeur de la distance nucléaire, appliquée par définition à un ensemble restreint de mots résilients, va de $0,1$ à $0,4$ environ.

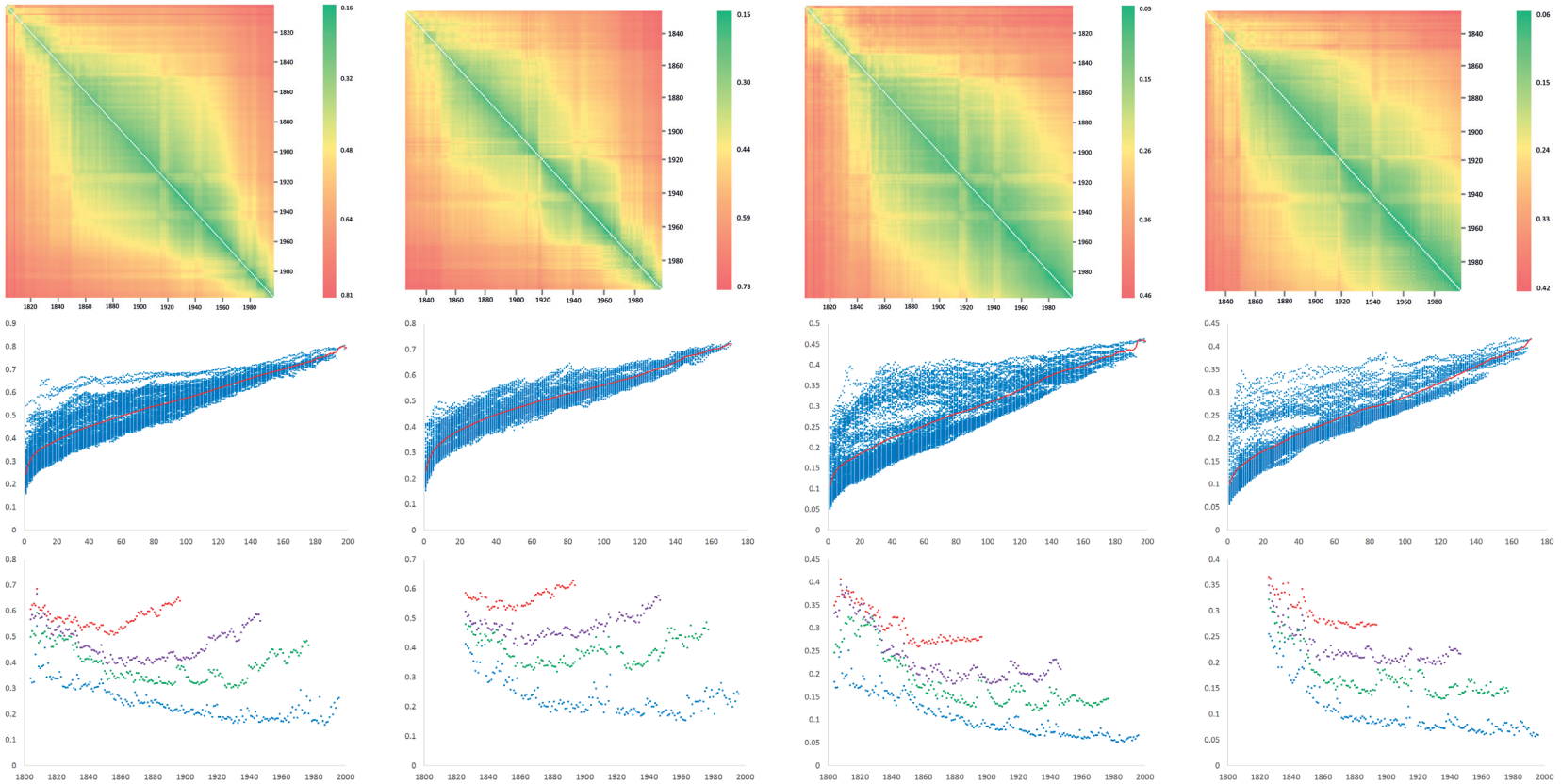


FIGURE 11.22 – **(1)** : Distances de Jaccard sur GDL ; **(2)** : Distances de Jaccard sur JDG ; **(3)** : Distances nucléaires sur GDL ; **(4)** : Distances nucléaires sur JDG ; **Haut** : Heatmap de la matrice des distances ; **Milieu** : Distances (bleu) et moyenne de ces distances (rouge) en fonction du nombre d'années de différence entre les sous-corpus ; **Bas** : Distances entre les années y_i et y_{i+n} avec $n = 1$ (bleu), $n = 20$ (vert), $n = 50$ (violet) and $n = 100$ (rouge)

Les deux distances partagent un comportement global similaire sur les deux corpus de JDG et GDL. Cependant, la distance nucléaire basée uniquement sur les mots les plus résilients, peut être éventuellement considérée comme une limite inférieure du changement linguistique montrant l'évolution des mots les plus stables.

Il est remarquable d'observer que les évolutions des distances avec le temps partage le même comportement malgré leurs différences. En effet, la distance de Jaccard appliquée au noyau résilient vaut zéro et n'utilise pas d'information concernant la fréquence des mots. La distance nucléaire quant à elle, utilise uniquement l'ordre fréquentiel d'un ensemble de mots très réduit et stable. Ces deux types différents d'informations convergent toutefois vers le même type de comportement pour les deux mesures.

Lors de la comparaison de la distance de Jaccard et de la distance nucléaire sur les Figures 11.6, 11.7, 11.19 et 11.20, nous observons que la distance nucléaire entre les sous-corpus implique que les années parmi les plus anciennes partagent les mêmes fluctuations qu'avec la distance de Jaccard tout en diminuant continuellement. Cependant, il est probable que cet effet soit dû à la faible quantité de données pour les années antérieures à 1850.

En général, la distance nucléaire semble diminuer lentement et en continu. Nous observons qu'il n'y a pas d'augmentation, mais plutôt une phase stable lorsque l'on considère deux sous-corpus séparés par plus de 20 ans. La distance nucléaire est également plus stable dans les dernières années. Afin de tester la robustesse de cette mesure, nous avons effectué une régression linéaire sur nos données et sur la période spécifiquement instable et considérée bruitée (1965-1998) pour les deux journaux.

Nous faisons l'hypothèse que la nature de l'évolution linguistique du corpus exclut les variations brutales et nous utilisons un modèle linéaire simple afin de tester le coefficient de régression avec les deux distances. La méthode donnant les résultats les plus stables devraient donc avoir un coefficient de régression plus élevé même avec un modèle aussi simple que la regression linéaire.

Les deux régressions sur l'ensemble des données pour les corpus de GDL et JDG sont représentées sur la Figure11.23. Nous observons que la distance nucléaire a un coefficient de régression plus élevé pour les deux corpus (0.8218 pour GDL et 0.6196 pour JDG) que la distance de Jaccard (0.6294 pour GDL et 0.3339 pour JDG).

Les régressions pour GDL et JDG sur la période particulièrement instable sont représentées sur la Figure 11.24. Nous observons également que la distance nucléaire a un coefficient de régression plus élevé sur cette courte période instable (0.2790 pour GDL et 0.1635 pour JDG) que la distance de Jaccard (0.0174 pour GDL et 0.00002 pour JDG).

Ces résultats suggèrent que, même si le bruit affecte toujours la distance nucléaire, elle est plus stable et robuste que celle de Jaccard.

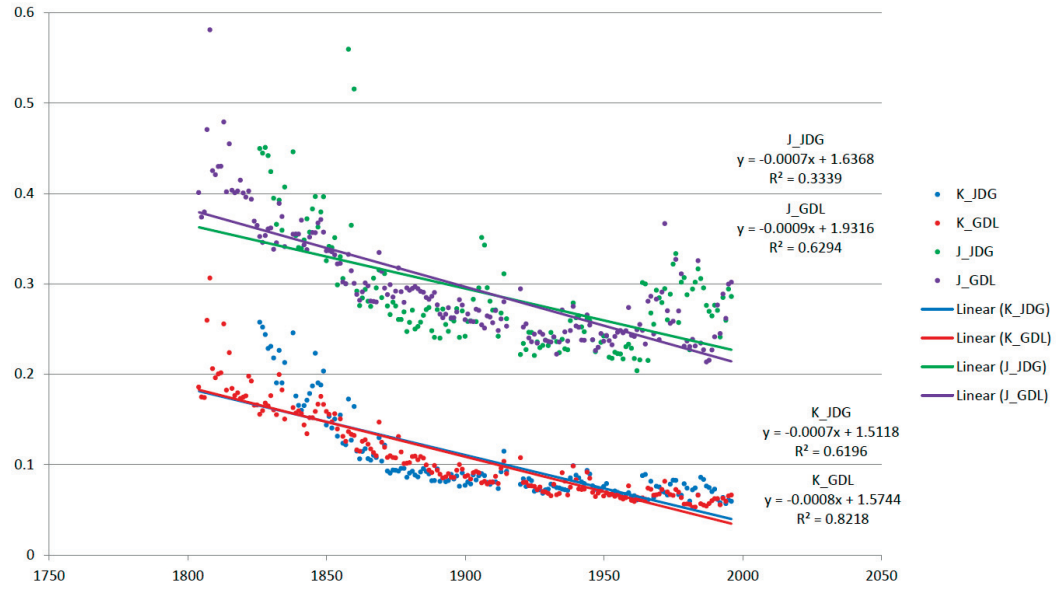


FIGURE 11.23 – Distances de Jaccard (violet pour le corpus de GDL et vert pour le corpus de JDG) et distances nucléaires (rouge pour GDL et bleu pour JDG) entre les années y_i et y_{i+1} en fonction des années y_i avec leur régression linéaire (lignes pleines) et le coefficient de régression

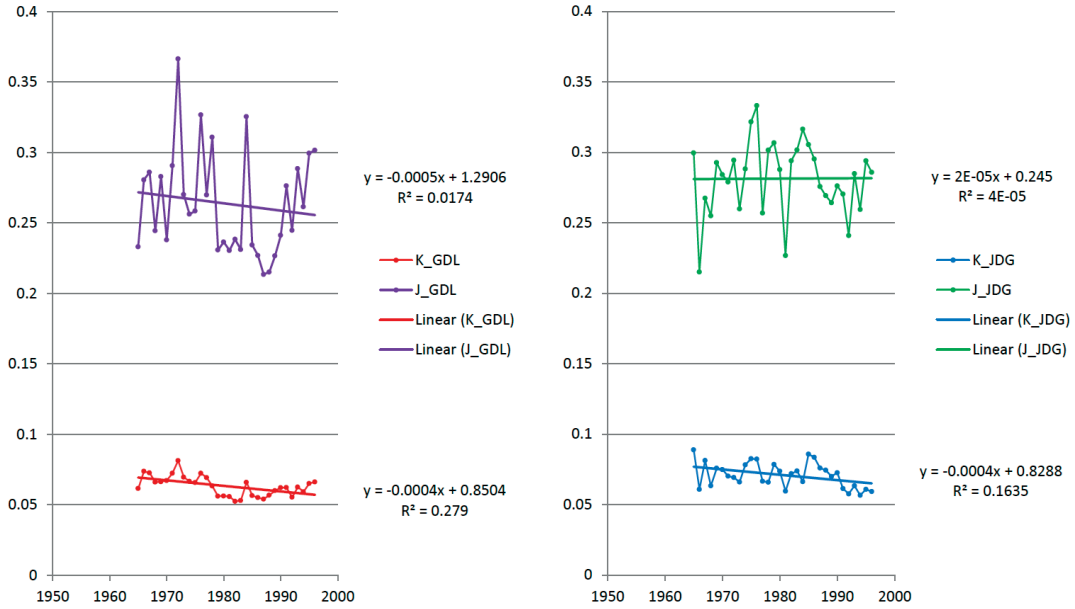


FIGURE 11.24 – Distances de Jaccard (violet pour le corpus de GDL et vert pour le corpus de JDG) et distances nucléaires (rouge pour GDL et bleu pour JDG) entre les années y_i et y_{i+1} en fonction des années y_i avec leur régression linéaire (lignes pleines) et le coefficient de régression sur la période de 1965 et plus

Simulations : tester la dépendance à la taille du corpus

Nous observons que la distance nucléaire est plus stable que la distance de Jaccard. Nous observons également que la distance nucléaire semble apporter une correction à la distance de Jaccard lors de la comparaison des sous-corpus dont la différence d'années est grande et dont la différence de taille est grande également puisque la taille du corpus augmente de façon quasi monotone avec le temps.

Cependant, seule la stabilité a pu être constatée et la question de l'indépendance par rapport à la taille des corpus ne peut pas être prouvée formellement. Nous avons donc conçu une expérience basée sur des simulations et nous calculons les distances de Jaccard et nucléaire sur des corpus simulés dans lesquels aucune évolution linguistique n'est inclue. Les corpus comparés auront un niveau de bruit linguistique qui dépend de la distribution générale observée des mots dans l'utilisation d'une langue, c'est-à-dire une distribution de type zipfienne et ce bruit dépend donc de la taille des sous-corpus qui sont considérés comme des échantillons représentatifs de la langue. Dans l'expérience de simulation, ces tailles sont fixées comme étant les mêmes que celles des corpus de GDL et JDG.

L'hypothèse de base est simple : chaque année nous générons un corpus sur la base d'une distribution de probabilité de type zipfienne, ensuite les distances sont calculées et comparées.

L'expérience se déroule donc en trois étapes distinctes :

1. Génération d'un lexique avec une distribution de fréquence.
2. Génération des mots (bag of words) pour chaque année en fonction du lexique généré à l'étape 1 et en fonction de la taille du corpus (ajustée sur les données de JDG ou GDL).
3. Calcul des distances Jaccard et nucléaire sur l'ensemble de mots générés et ce pour chaque paire d'années.

Nous concevons d'abord le processus de génération du lexique selon un paramètre unique qui est la taille fixée pour le lexique généré. La distribution de fréquence est alors calculée en fonction de la taille et la loi Zipf qui est une bonne approximation du comportement de la distribution de fréquence des mots dans la langue. Le code python pour la fonction de génération du lexique est donné dans les lignes suivantes :

```
def generate_lexique(size):
    lex=[]
    freq=[]
    freq_tot=0
    for i in range(size):
        f=1.0/(i+1)
        freq_tot+=f
    for i in range(size):
        f=1.0/(freq_tot*(i+1))
        lex.append(str(i+1))
        freq.append(f)
    return [lex,freq]
```

Cette fonction génère deux listes. La première est une liste de chaînes de caractères correspondant aux mots du lexique généré. Ces mots sont en réalité des nombres transformés en chaînes de caractères représentant des identifiants uniques. La seconde liste correspond à la distribution de fréquence des mots générés, dans cet exemple la distribution prend la forme d'une loi de Zipf exacte. Il est également possible d'utiliser une loi de Heaps moyennant une estimation de ses paramètres.

Une fois que le lexique est généré, nous construisons une autre fonction afin de générer une liste de mots selon le nombre de mots désirés comme paramètre. Le code python pour la génération d'ensembles de mots est donné dans les lignes suivantes :

```
import random

def generate_ensemble(size,lex):
    x=[]
    for i in range(size):
        f_rand=random.uniform(0.0,1.0)
        f=0.0
        found=0
        compteur=0
        while found==0:
            f+=lex[1][compteur]
            if f_rand<f:
                x.append(lex[0][compteur])
                found=1
            compteur+=1
    return x
```

Le processus commence donc par la génération d'un lexique disposant d'une taille fixe donnée, puis la génération des corpus annuels selon le nombre exact de mots pour chaque sous-corpus annuel réel de JDG et GDL. Pour terminer, il suffit d'appliquer la distance de Jaccard et la distance nucléaire sur ces ensembles afin d'observer le comportement des distances appliquées à des données sans aucune évolution linguistique.

Ce processus peut être répété pour différentes tailles de lexique. Les tailles des échantillons générés sur la base du lexique sont connues et fixées égales aux tailles réelles des sous-corpus annuels de JDG et GDL. La taille du lexique de base est un paramètre inconnu, difficile à estimer et susceptible de changer avec le temps.

Nous utilisons donc les observations des corpus GDL et JDG pour extraire plusieurs valeurs différentes et réalistes de la taille du lexique que nous devrions générer afin de couvrir différents cas extrêmes. Nous avons fixé les tailles de lexiques générés aux valeurs suivantes : 30 000, 50 000, 100 000, 200 000, 500 000, 1 000 000, 2 000 000 et 5 000 000.

Les résultats de simulation pour chaque taille de lexique sont présentés dans les Figures 11.25 et 11.26 pour la distance de Jaccard et les Figures 11.27 et 11.28 pour la distance nucléaire.

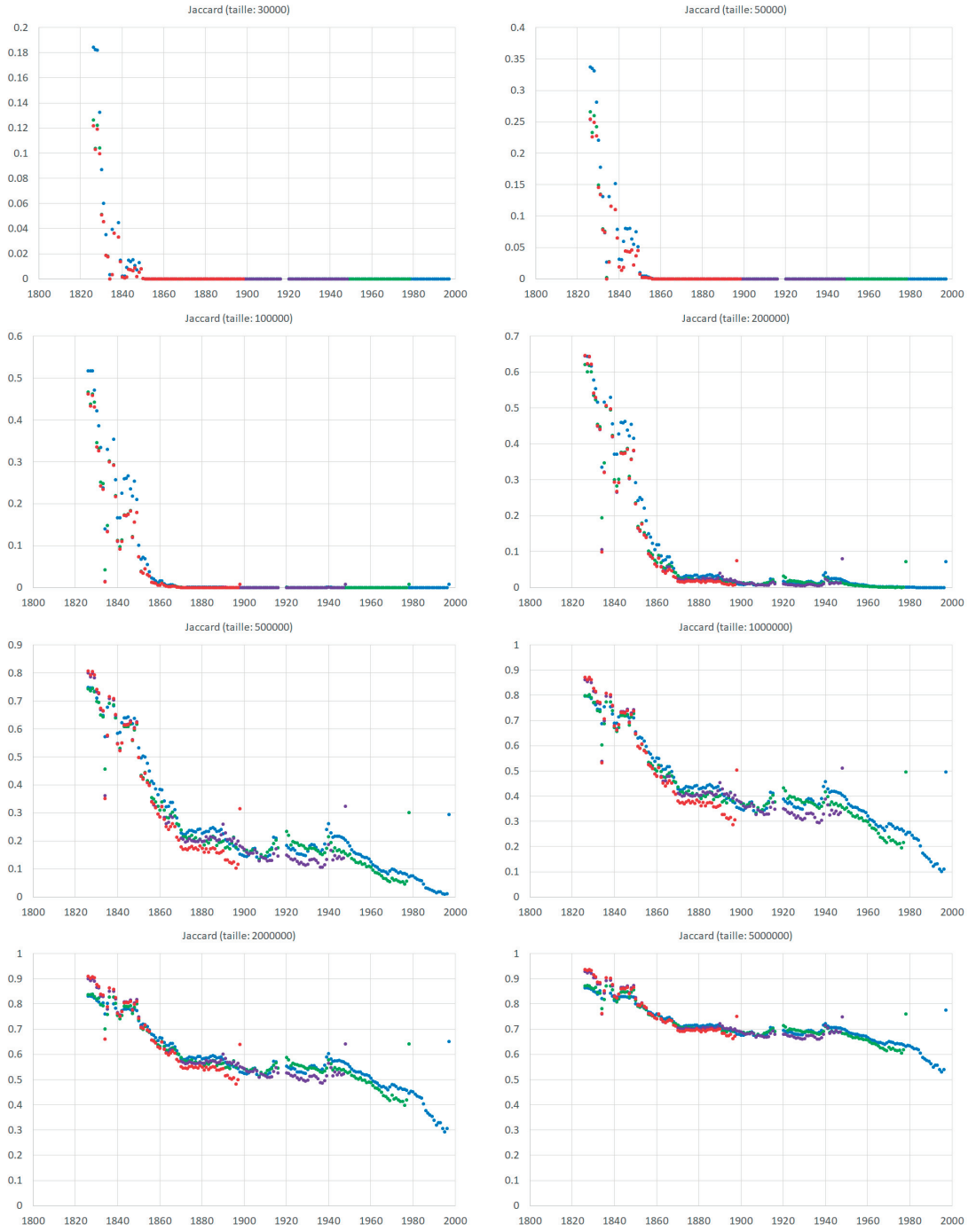


FIGURE 11.25 – Distances de Jaccard entre les années y_i et y_{i+n} avec $n = 1$ (bleu), $n = 20$ (vert), $n = 50$ (violet) et $n = 100$ (Rouge) pour les tailles de corpus de JDG et la taille de lexique simulée égale à 30 000 (première ligne, gauche), 50 000 (première ligne, droite), 100 000 (deuxième ligne, gauche), 200 000 (deuxième ligne , droite), 500 000 (troisième ligne, gauche), 1 000 000 (troisième ligne, droite), 2 000 000 (quatrième ligne, gauche) et 5 000 000 (quatrième ligne, droite)

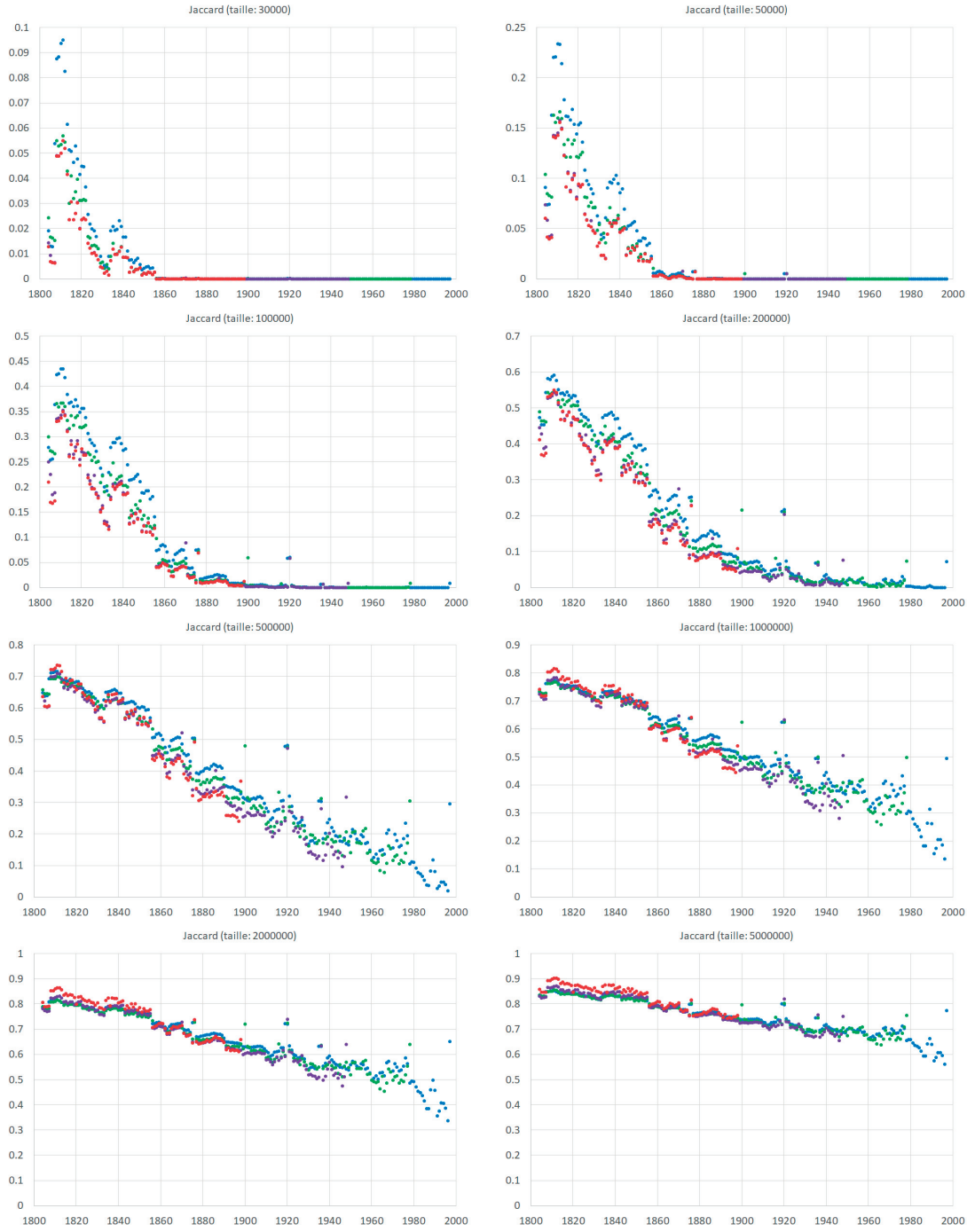


FIGURE 11.26 – Distances de Jaccard entre les années y_i et y_{i+n} avec $n = 1$ (bleu), $n = 20$ (vert), $n = 50$ (violet) et $n = 100$ (Rouge) pour les tailles de corpus de GDL et la taille de lexique simulée égale à 30 000 (première ligne, gauche), 50 000 (première ligne, droite), 100 000 (deuxième ligne, gauche), 200 000 (deuxième ligne , droite), 500 000 (troisième ligne, gauche), 1 000 000 (troisième ligne, droite), 2 000 000 (quatrième ligne, gauche) et 5 000 000 (quatrième ligne, droite)

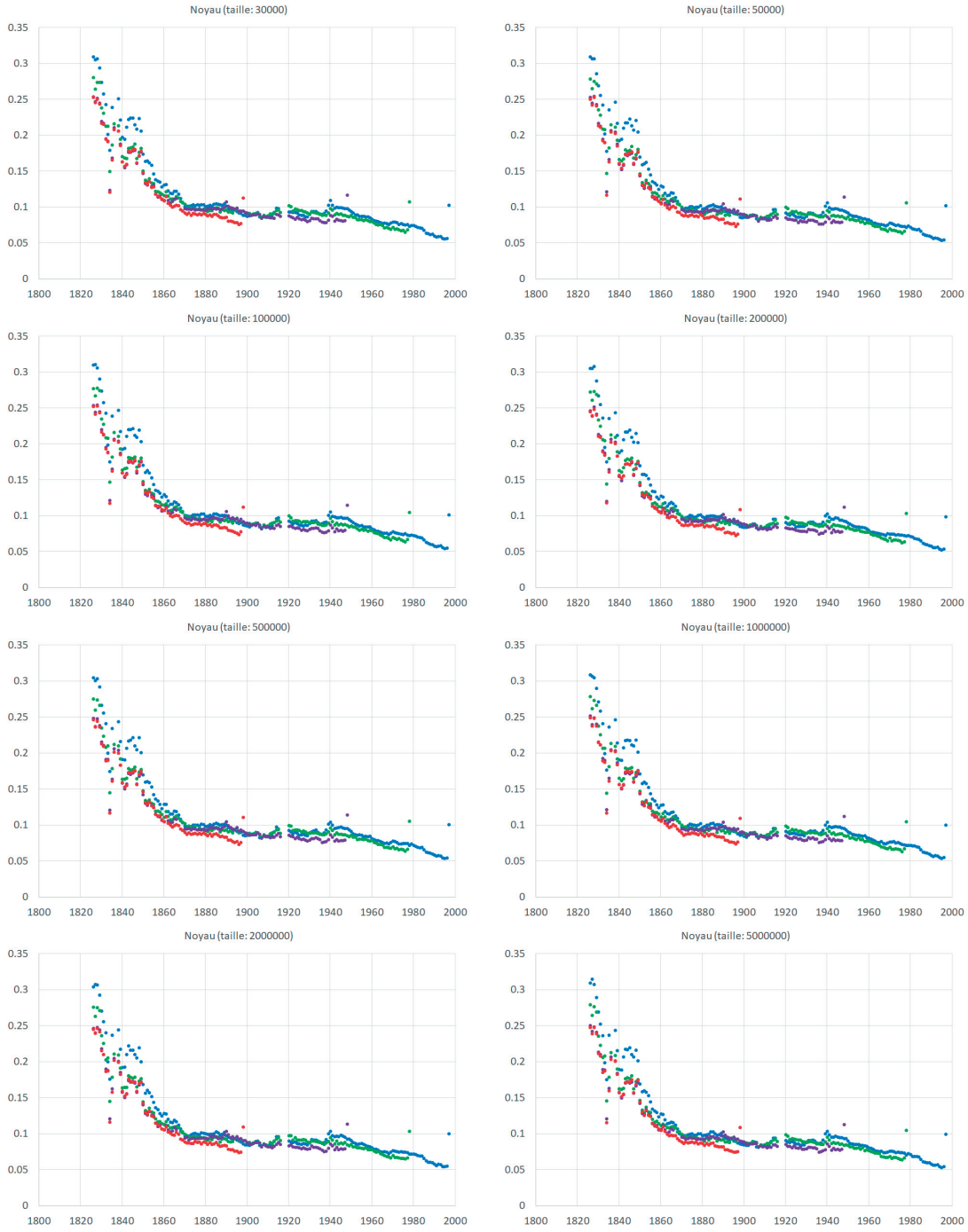


FIGURE 11.27 – Distances nucléaires entre les années y_i et y_{i+n} avec $n = 1$ (bleu), $n = 20$ (vert), $n = 50$ (violet) et $n = 100$ (Rouge) pour les tailles de corpus de JDG et la taille de lexique simulée égale à 30 000 (première ligne, gauche), 50 000 (première ligne, droite), 100 000 (deuxième ligne, gauche), 200 000 (deuxième ligne , droite), 500 000 (troisième ligne, gauche), 1 000 000 (troisième ligne, droite), 2 000 000 (quatrième ligne, gauche) et 5 000 000 (quatrième ligne, droite)

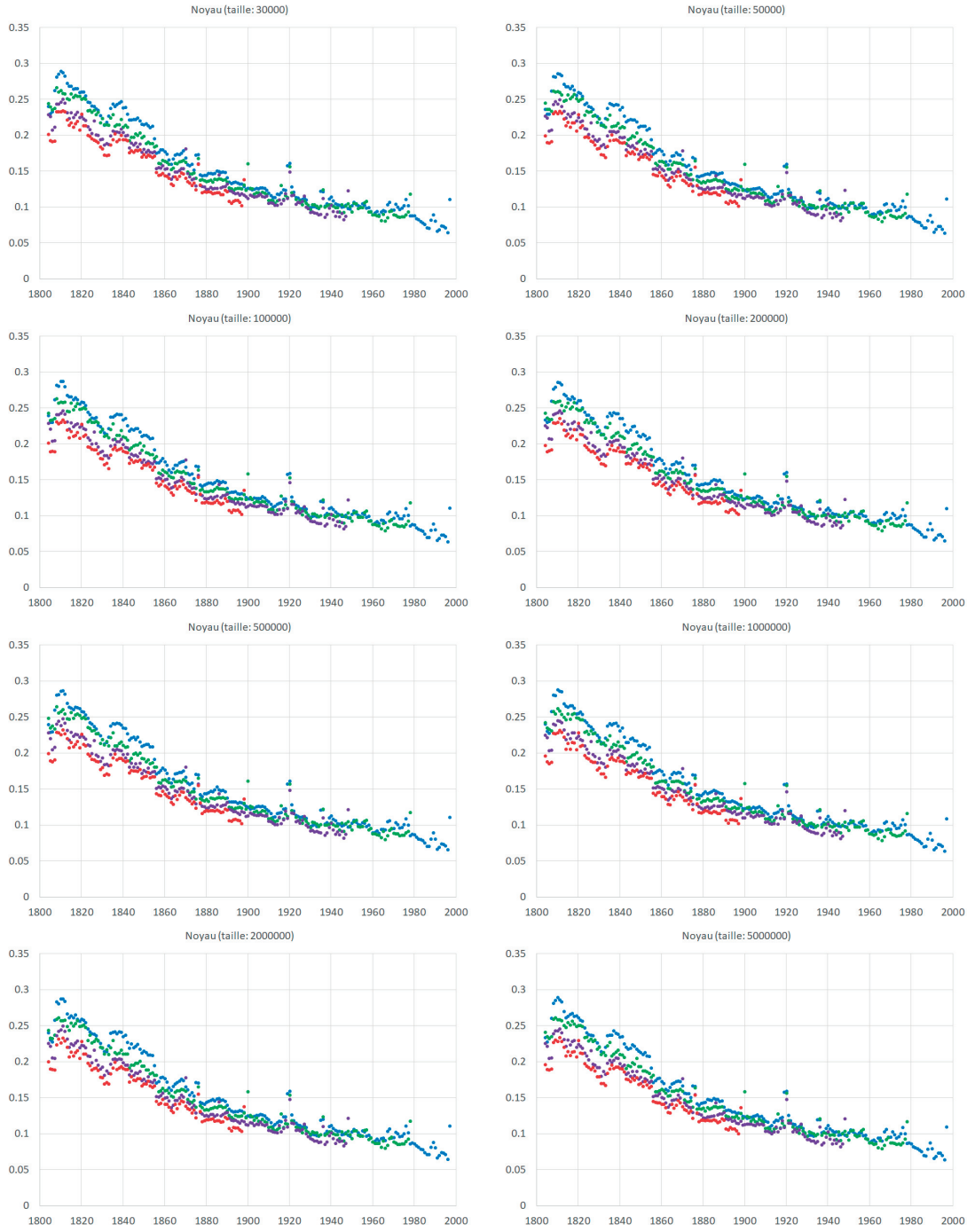


FIGURE 11.28 – Distances nucléaires entre les années y_i et y_{i+n} avec $n = 1$ (bleu), $n = 20$ (vert), $n = 50$ (violet) et $n = 100$ (Rouge) pour les tailles de corpus de GDL et la taille de lexique simulée égale à 30 000 (première ligne, gauche), 50 000 (première ligne, droite), 100 000 (deuxième ligne, gauche), 200 000 (deuxième ligne , droite), 500 000 (troisième ligne, gauche), 1 000 000 (troisième ligne, droite), 2 000 000 (quatrième ligne, gauche) et 5 000 000 (quatrième ligne, droite)

Nous observons dans les Figures 11.25 et 11.26 que la distance de Jaccard est une "bonne" distance linguistique lorsque la taille du lexique est faible. En effet, nous attendons un résultat de zéro sur l'ensemble des distances, car aucun modèle d'évolution linguistique n'a été ajouté à la simulation. Le fait que la distance de Jaccard soit proche de zéro signifie que le corpus considéré comme un échantillon du lexique du langage réel est représentatif. De façon plus surprenante, nous observons dans les Figures 11.27 et 11.28 que la distance nucléaire est une distance linguistique non idéale dans tous les cas de simulation sans exception. En effet, même dans le cas d'une taille faible du lexique, et donc d'un échantillon représentatif, la distance nucléaire appliquée sur l'échantillon montre clairement une distance non négative alors qu'elle devrait être égale à zéro. La distance nucléaire est donc sensible à la variation de taille des subcorpus étudiés.

Cependant, un avantage évident de la distance est visible sur les Figures 11.27 et 11.28, la distance est extrêmement stable. Elle semble être totalement indépendante de la taille du lexique linguistique qui est le paramètre que nous ne connaissons pas et ne dépend que des tailles annuelles des sous-corpus étudiés. Grossièrement, nous pourrions dire que la distance est fausse, mais par contre elle est stable. Elle est donc plus avantageuse, car dans le cas de la distance nucléaire, il devient possible d'estimer la distance nucléaire générée par le bruit linguistique sans aucune évolution et de la comparer à la distance nucléaire appliquée sur le corpus empirique qui contient du bruit et en même temps l'évolution linguistique.

Bien que rien ne prouve que l'effet dû à la variation de taille des sous-corpus et celui dû à l'évolution linguistique réel du corpus soient additifs, nous pouvons fournir une estimation approximative de ces effets en supposant que la distance calculée sur les corpus de JDG ou GDL est égale à la distance nucléaire linguistique plus la distance nucléaire du bruit dû aux variations de taille du corpus. Sous cette hypothèse, nous pouvons tenter de quantifier le changement linguistique du corpus en soustrayant les valeurs de distance nucléaire simulées des valeurs de distance nucléaire calculées empiriquement. Les résultats sont présentés dans les Figures 11.29 et 11.30.

Enfin, nous observons également dans les Figures 11.25, 11.26, 11.27 et 11.28 que toutes les courbes se chevauchent, ce qui n'est le cas pour aucune des distances appliquées sur les données réelles des corpus GDL et JDG. Cela suggère que la comparaison différentielle des courbes peut permettre la quantification du changement linguistique du corpus malgré l'évolution de la taille de celui-ci. Plus simplement, cela suggère l'existence du changement linguistique malgré la dépendance de la mesure initiale à l'évolution de la taille du corpus. Effectivement, dans le cas des données réelles des corpus de JDG et GDL, nous avons observé que la courbe ne se chevauche pas et donc que l'évolution linguistique est différente de zéro.

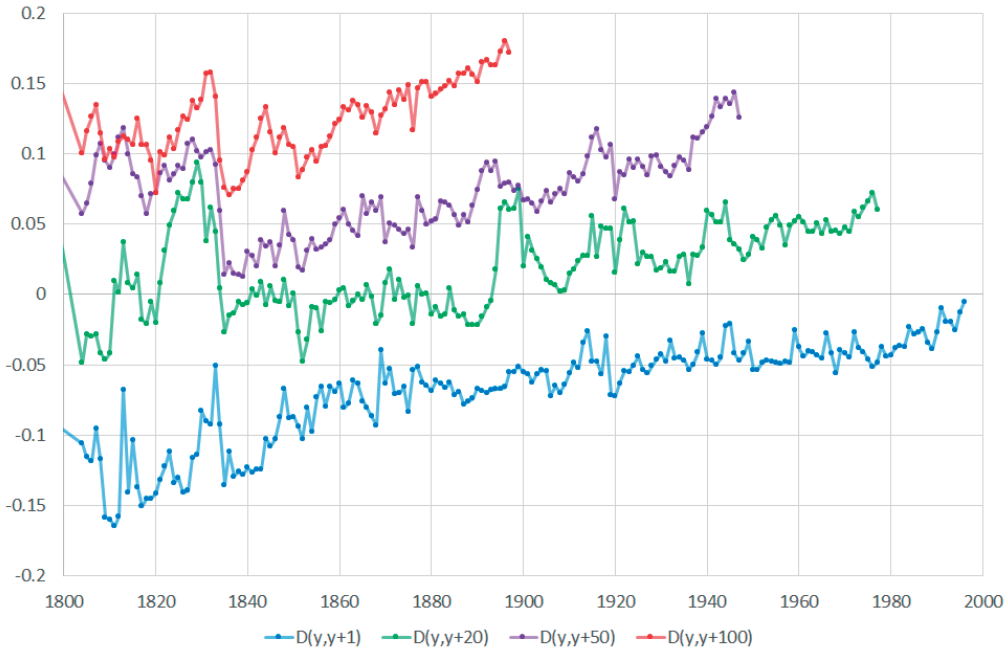


FIGURE 11.29 – Différence entre les distances nucléaires réelles et les distances nucléaires simulées sur le corpus de GDL

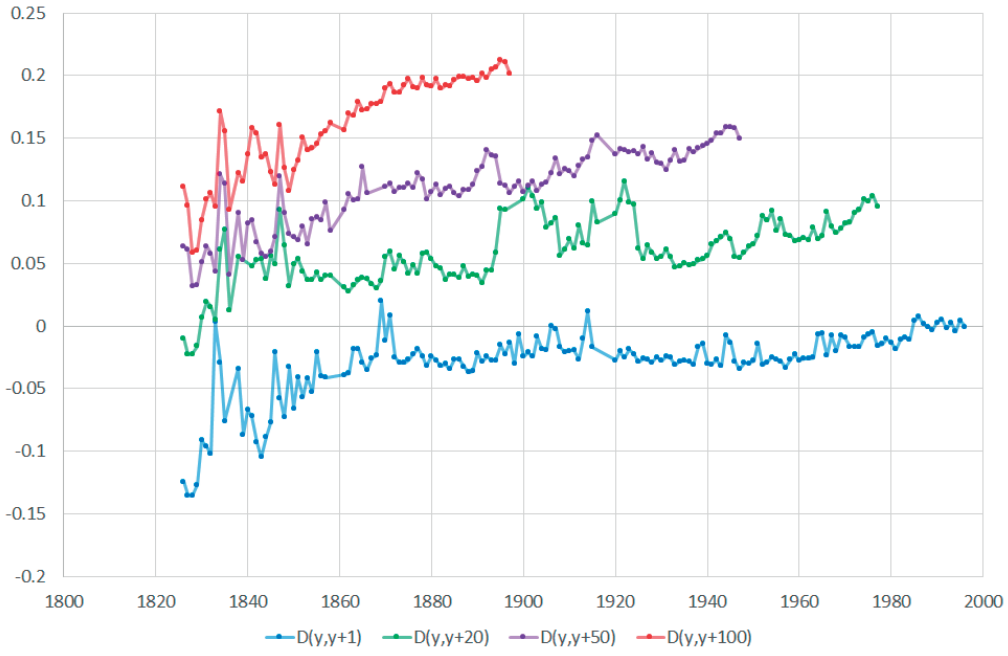


FIGURE 11.30 – Différence entre les distances nucléaires réelles et les distances nucléaires simulées sur le corpus de JDG

Synthèse de l'analyse de distances

Deux définitions de distance différentes ont été appliquées aux corpus de GDL et JDG dans le but de quantifier leurs changements linguistiques. Nous avons d'abord utilisé la distance de Jaccard sur l'ensemble du corpus avec un filtre fréquentiel.

Nos observations des Figures 11.2 à 11.7 soutiennent l'hypothèse de l'existence de changements linguistiques quantifiables sur la base uniquement des mots, même si nous observons aussi que la distance de Jaccard est potentiellement sensible au bruit. En outre, la distance de Jaccard est connue pour être également sensible aux fluctuations de taille du corpus (Muller, 1980) (Brunet, 2003), nous avons donc utilisé le concept de résilience et de noyau résilient afin d'étudier la partie la plus stable de la langue au travers de ces corpus.

Nous avons ensuite utilisé une distance nucléaire basée sur la comparaison du rang fréquentiel des mots du noyau. De manière surprenante, les Figures 11.17, 11.18, 11.19 et 11.20 nous permettent d'observer un comportement similaire à la distance de Jaccard sur l'ensemble du corpus. Cette observation soutient l'hypothèse selon laquelle les informations sur les distances linguistiques extraites par contrôle de la présence et l'absence des mots sur l'ensemble du corpus peuvent être aussi extraites en utilisant un ensemble réduit de mots résilients du noyau et en leur appliquant la distance nucléaire.

De plus, la distance nucléaire a une sensibilité réduite au bruit, limitant l'effet de la "contamination" des années contenant un taux de bruit plus élevés comme la période de 1900 à 1915 pour JDG et la période de 1965 et plus, pour les deux journaux. Les observations pour cette distance, comme pour celle de Jaccard, montrent une diminution durant la période antérieure à 1870 qui est une période faiblement représentative de la langue en raison de la petite taille du corpus. Après cette période, les distances d'une année à l'autre semblent diminuer lentement, mais avec plus de stabilité. En outre, les distances d'une année à 20, 50 ou 100 ans plus tard restent stables.

A partir de nos expériences sur les deux corpus de GDL et JDG, nous avons fait une série d'observations qui soutiennent l'existence d'un changement linguistique continu et relativement constant dans ces corpus. Nous avons essayé plusieurs méthodes pour quantifier l'évolution de ce changement tout en évitant d'incorporer du bruit, de la fluctuation de la taille du corpus, la variation des sujets d'articles ou la qualité de l'OCR. L'objectif de la notion de noyau résilient étant de cibler autant que possible la langue au travers du corpus.

Si ces mesures montrent une manière de quantifier le changement linguistique, nous n'avons pas d'indication d'une accélération ou d'une décélération de l'évolution de ce changement sur les périodes de 1804-1997 (GDL) et 1826-1997 (JDG). Cependant, ces méthodes devraient être appliquées sur d'autres corpus dont ceux dont les données sont disponibles après 1997. Cela permettrait de vérifier si cette stabilité observée est maintenue pendant la période 1998-2017, où beaucoup de technologies interviennent directement sur notre expression textuelle et donc notre langue, accélérant potentiellement l'évolution linguistique (Kaplan, 2014).

Même si nous avons observé que la distance nucléaire est moins sensible au bruit du corpus, il reste difficile de prouver l'indépendance de cette distance par rapport à des variables comme l'évolution de la taille du corpus. Pour ce faire, nous avons effectué des simulations des distances de Jaccard et nucléaires en générant un ensemble aléatoire de mots de même taille que les corpus observés de JDG et GDL, basé sur la distribution de Zipf, mais sans aucune évolution linguistique. Nous avons également observé la stabilité particulière de la distance nucléaire par rapport à l'effet de l'échantillonnage, ce qui nous a permis d'estimer l'effet du bruit selon la variation de la taille du corpus. Nous avons ensuite proposé d'estimer les changements linguistiques du corpus en soustrayant les distances simulées (représentant l'effet de la taille du corpus) des distances calculées sur les données réelles.

De grandes bases de données de journaux scannés ouvrent de nouvelles voies pour étudier l'évolution de la langue (Westin et Geisler, 2002) (Fries et Lehmann, 2006) (Bamford *et al.*, 2013). Cependant, ces études devraient être menées avec des méthodologies solides afin d'éviter une interprétation erronée des résultats liés à la variation de taille des sous-ensembles ou à des généralisations insuffisamment motivées de ces résultats, obtenus à partir de corpus spécifiques de presse, à l'évolution de la langue en général.

Dans cette section, nous avons introduit la notion de noyau résilient comme une approche possible pour étudier les changements linguistiques du corpus sous l'angle de l'étude des mots les plus stables. En effet, mettre l'accent sur des mots stables et leur distribution relative est susceptible de rendre les interprétations plus robustes. Les résultats ont été calculés à partir de deux corpus indépendants. Il est frappant de voir que la plupart des résultats obtenus pour chacun d'entre eux sont semblables. Les compositions des noyaux en termes de parties du discours sont également similaires.

La distance nucléaire, appliquée aux mots du noyau résilient pour mesurer les changements linguistiques, s'est révélée robuste pour les cas d'erreur d'OCR et de bruit. En outre, nous avons observé que l'étude des mots du noyau par la distance nucléaire permet d'extraire des informations semblables à celles de la distance de Jaccard appliquée à l'ensemble du corpus. Cela suggère que nos méthodes mesurent effectivement des phénomènes linguistiques généraux au-delà de la spécificité des corpus choisis pour cette étude.

11.2 Entropie

L'entropie est une notion de thermodynamique des sciences physiques quantifiant le désordre ou la désorganisation d'un système. La notion d'entropie s'est généralisée dans les domaines des mathématiques et de l'informatique, elle correspond intuitivement à la quantité d'informations contenues dans un système. Nous utilisons la formule d'entropie de Shannon suivante :

$$H = -\sum_{i=1}^{N} f_i \; ln(f_i)$$

Où f_i représente la fréquence d'apparition du mot i dans le sous-corpus considéré. Nous quantifions alors l'information annuelle contenue dans GDL et JDG afin d'en mesurer l'évolution. Les mesures de l'entropie de Shannon sur l'ensemble des mots des sous-corpus annuels de GDL et JDG sont présentés dans la Figure 11.31.

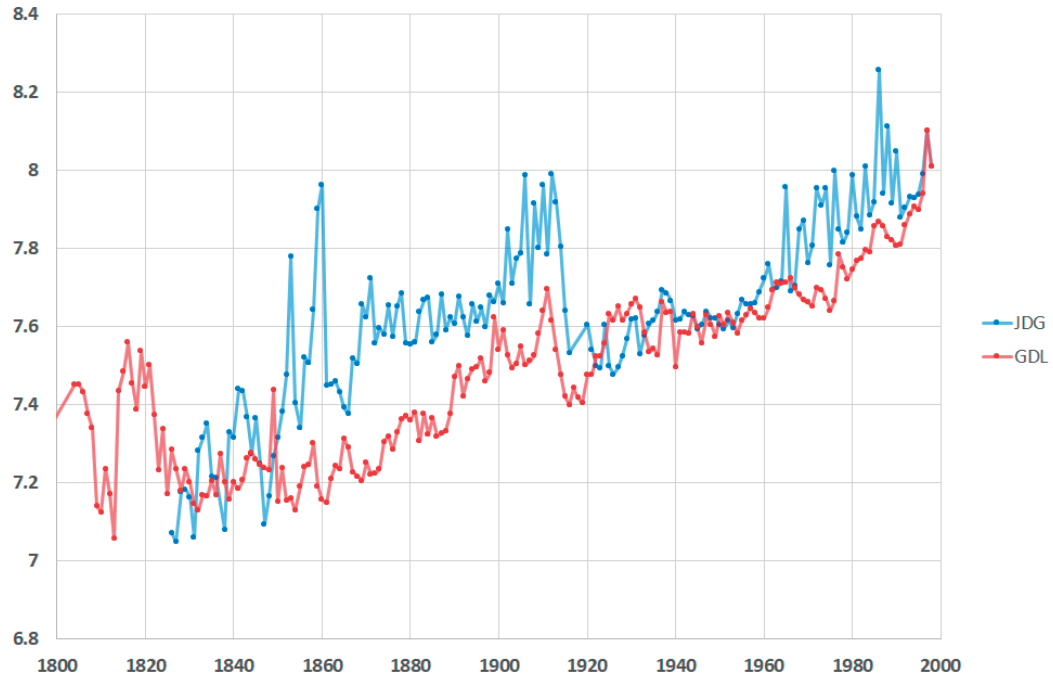


FIGURE 11.31 – Entropie de Shannon des 1-grammes de GDL (rouge) et JDG (bleu)

Nous observons que l'entropie des deux journaux à tendance à augmenter. Nous constatons divers régimes comme la période 1850-1920 où l'entropie de JDG surpasse de façon importante celle de GDL, la période 1920-1965 où l'entropie des deux journaux a tendance à coïncider et la période 1965 et plus où l'entropie de JDG subit des variations importantes tout en étant plus élevée que celle de GDL. La période 1830 et moins pour GDL semble être aussi sujette à des variations importantes potentiellement causées par la faible taille de corpus pour les années les plus anciennes.

En outre, nous observons également une réduction brutale de l'entropie pour les deux corpus aux alentours du début de la première guerre mondiale.

Il est clair que le niveau d'informations contenues dans un corpus n'est pas indépendant de sa taille, car si celui-ci augmente, alors la diversité lexicale aura tendance à augmenter aussi. Nous avons la possibilité de construire une mesure d'entropie moins dépendante de la taille du corpus via la notion de noyau résilient.

Il s'agit de renormaliser toute les probabilités du corpus en ne tenant compte que de l'apparition des mots du noyau résilient du corpus. Toutefois, afin de vérifier sous quelles conditions ces mesures d'entropie sont indépendantes de la taille du corpus, nous avons effectué des simulations produites à l'aide du même processus décrit pour les simulations de distance nucléaire. Comme pour celles-ci, nous avons considéré les tailles de vocabulaire suivantes : 30 000, 50 000, 100 000, 200 000, 500 000, 1 000 000, 2 000 000 et 5 000 000, couvrant une grande diversité de situation et notamment celles rencontrées dans les deux corpus de JDG et GDL.

Nous obtenons une série de graphiques représentant l'évolution de la mesure au cours du temps pour chaque taille de lexique considéré. Les mesures d'entropie obtenues pour l'ensemble de JDG sont présentées dans la Figure 11.32 et pour son noyau dans la Figure 11.33. Les mesures d'entropie obtenues pour l'ensemble de GDL sont présentées dans la Figure 11.34 et pour son noyau dans la Figure 11.35.

Une fenêtre de valeur encadrant la médiane est utilisée afin de vérifier la stabilité de la mesure et son indépendance vis-à-vis de la taille du corpus. La valeur choisie est de 0.01 en mesure d'entropie. Nous constatons une évidente évolution de la mesure d'entropie totale avec la taille du corpus. Toutefois, la mesure d'entropie limitée aux éléments du noyau résilient ne varie pas en fonction de la taille du corpus, si ce n'est dans les toutes premières années quand la taille du corpus par année est faible et probablement non représentative.

La mesure d'entropie restreinte au noyau résilient est donc suffisamment stable et indépendante de la taille du corpus pour les années postérieures à 1840. Nous avons vérifié si une distance simple de type cosinus appliquée sur les vecteurs annuels composés de toutes les contributions entropiques de chaque mot est également indépendante de l'évolution de la taille du corpus et avons conclu que la dépendance à la taille du corpus se retrouve dans ces distances. Ce n'est donc qu'en agrégeant les contributions entropiques de tous les éléments que l'on obtient une mesure indépendante de la taille du corpus sur chaque simulation.

Nous avons donc calculé la mesure d'entropie basée sur les mots composant le noyau résilient afin comparer l'évolution diachronique des corpus de JDG et GDL. Toutefois, pour comparer formellement les deux journaux, nous utilisons comme ensemble résilient l'intersection des noyaux de GDL et JDG, nous renormalisons les fréquences d'apparition sur cette partie du corpus et nous calculons l'entropie de Shannon sur cet ensemble de mots résilients et commun aux deux journaux. Les mesures d'entropie de Shannon sur l'intersection des noyaux résilients des corpus de GDL et JDG sont présentés dans la Figure 11.36.

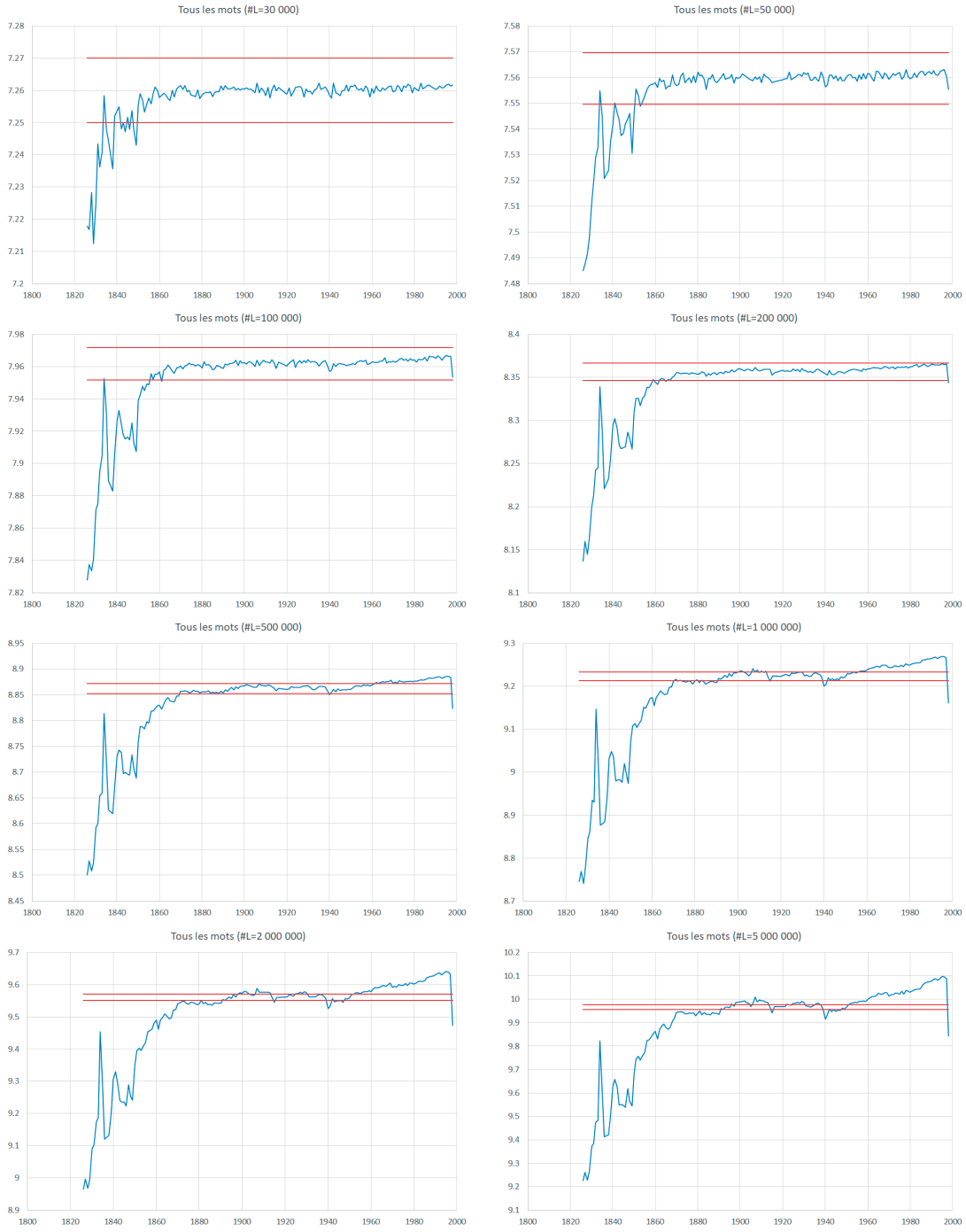


FIGURE 11.32 – Entropie de chaque année pour les tailles de corpus de JDG et la taille de lexique simulée égale à 30 000 (première ligne, gauche), 50 000 (première ligne, droite), 100 000 (deuxième ligne, gauche), 200 000 (deuxième ligne , droite), 500 000 (troisième ligne, gauche), 1 000 000 (troisième ligne, droite), 2 000 000 (quatrième ligne, gauche) et 5 000 000 (quatrième ligne, droite)

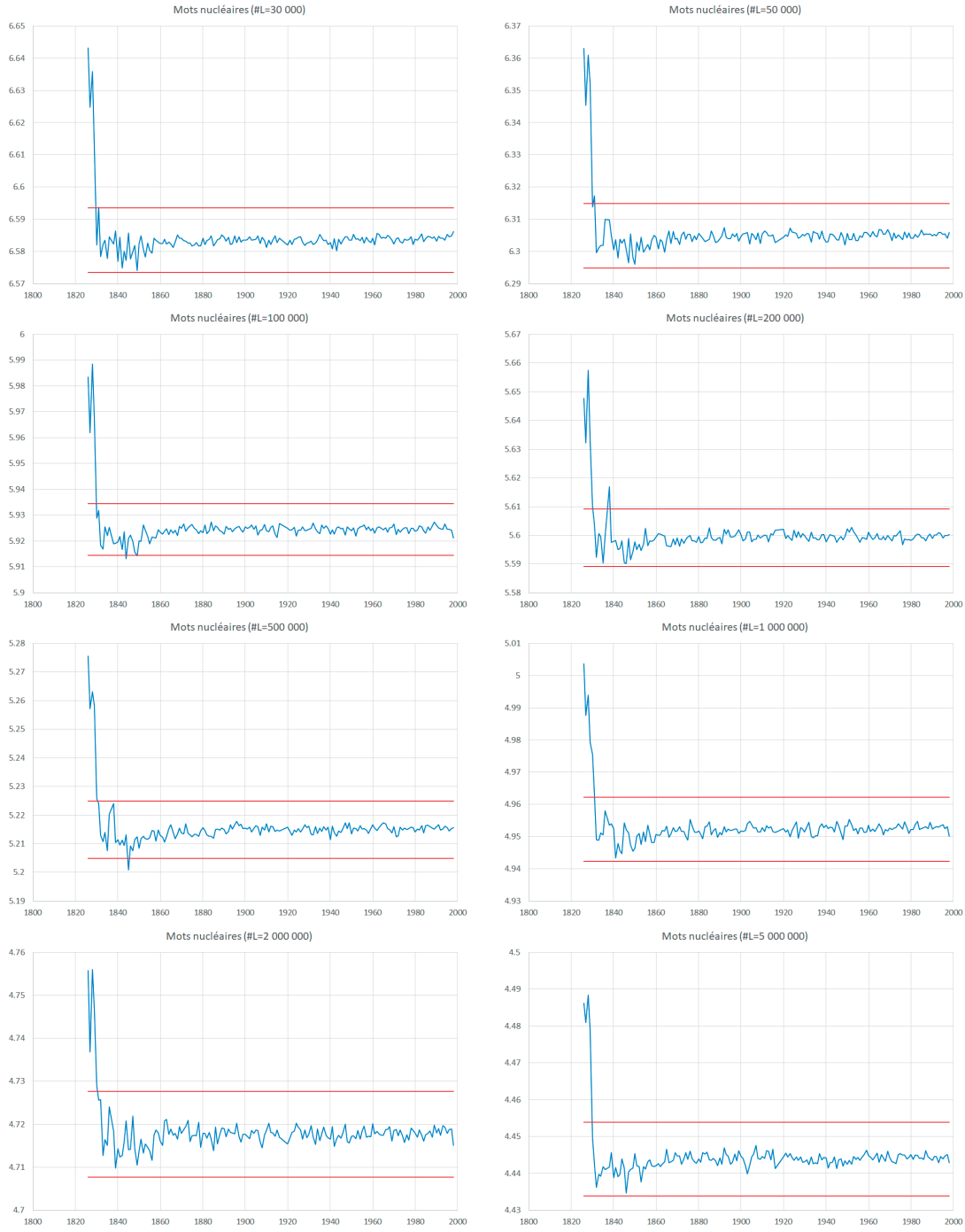


FIGURE 11.33 – Entropie nucléaire de chaque année pour les tailles de corpus de JDG et la taille de lexique simulée égale à 30 000 (première ligne, gauche), 50 000 (première ligne, droite), 100 000 (deuxième ligne, gauche), 200 000 (deuxième ligne , droite), 500 000 (troisième ligne, gauche), 1 000 000 (troisième ligne, droite), 2 000 000 (quatrième ligne, gauche) et 5 000 000 (quatrième ligne, droite)

FIGURE 11.34 – Entropie de chaque année pour les tailles de corpus de GDL et la taille de lexique simulée égale à 30 000 (première ligne, gauche), 50 000 (première ligne, droite), 100 000 (deuxième ligne, gauche), 200 000 (deuxième ligne , droite), 500 000 (troisième ligne, gauche), 1 000 000 (troisième ligne, droite), 2 000 000 (quatrième ligne, gauche) et 5 000 000 (quatrième ligne, droite)

FIGURE 11.35 – Entropie nucléaire de chaque année pour les tailles de corpus de GDL et la taille de lexique simulée égale à 30 000 (première ligne, gauche), 50 000 (première ligne, droite), 100 000 (deuxième ligne, gauche), 200 000 (deuxième ligne , droite), 500 000 (troisième ligne, gauche), 1 000 000 (troisième ligne, droite), 2 000 000 (quatrième ligne, gauche) et 5 000 000 (quatrième ligne, droite)

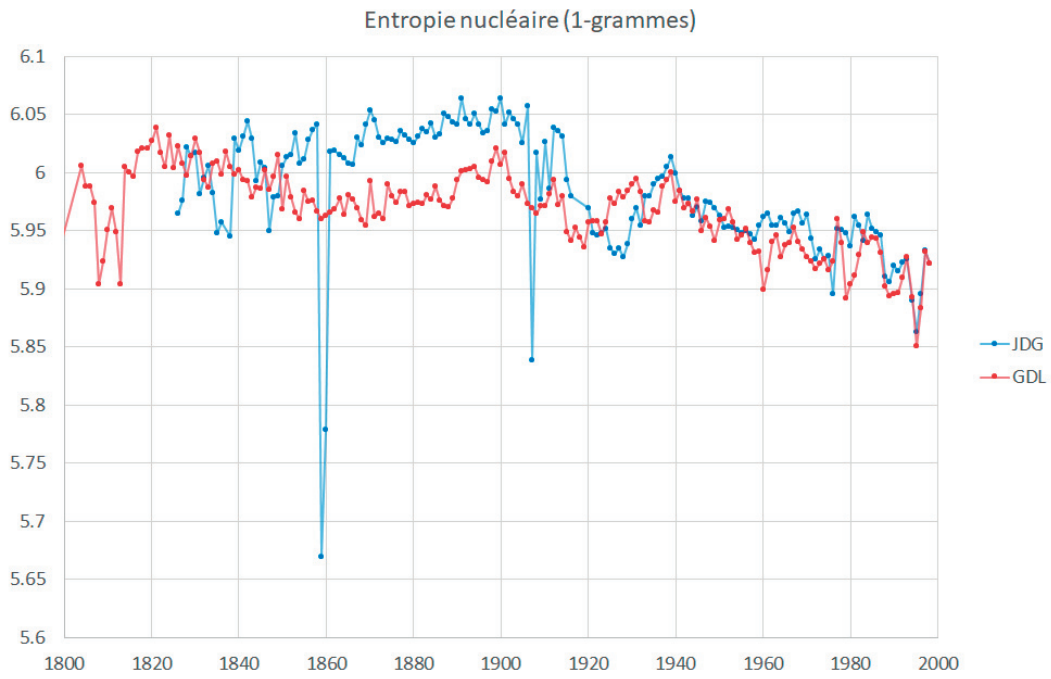


FIGURE 11.36 – Entropie de Shannon des 1-grammes de GDL (rouge) et JDG (bleu) calculée sur la fréquence renormalisée des mots communs aux noyaux résilients de GDL et JDG

Nous observons ainsi l'évolution de l'entropie des deux journaux sur une base qui permet la comparaison. Nous constatons cette fois une tendance à la diminution de l'entropie avec le temps. Trois années présentent une valeur extrême pour le journal JDG, il s'agit de 1859, 1860 et 1907. Les valeurs extrêmes sont caractéristiques d'une diminution forte de la qualité et la résolution des pages scannées, impactant directement la reconnaissance OCR sur ces années, elles doivent donc être retirées de la plupart des analyses effectuées. Il est frappant de constater que la différence d'entropie entre les deux journaux est élevée dans la période 1850-1920. Toutefois, comme pour l'étude de la distance nucléaire, l'étude du noyau résilient permet de stabiliser les variations importantes initialement constatées pour la période 1965 et plus dans le corpus de JDG. Sur ce noyau, l'impact des deux guerres mondiales reste important, mais semble par contre se résorber avec le temps. Enfin une tendance à une faible augmentation est constatée dans la période d'avant-guerre et ensuite une tendance à la diminution dans la période après-guerre jusque l'année 1998.

Dans le travail (Juola, 2013), l'entropie est utilisée afin de mesurer la complexité de la culture dans le corpus de Google Books. La même mesure montre une augmentation sur les corpus de GDL et JDG ainsi que via des simulations dépourvues d'évolution lexicale. L'entropie nucléaire nous permet de mesurer l'entropie du noyau de façon indépendante (dès 1840) par rapport à l'évolution de la taille du corpus. Nous observons ainsi une diminution de l'entropie nucléaire dès 1940 contrairement à l'augmentation de l'entropie constatée sur l'entier du corpus.

11.3 Chronocloud

Chronoclouds classiques

Dans cette section, nous analysons les chronoclouds de mots des corpus de JDG et GDL. Ces chronoclouds sont calculés en haute résolution afin de permettre le zoom et donc la découverte d'autre mots particuliers du corpus. Toutefois, nos analyses concerneront principalement les mots les plus visibles sur le chronocloud représenté à l'échelle d'une feuille au format classique A4. Nous pouvons imaginer d'autres échelles ou interfaces de représentation permettant de découvrir des mots moins fréquents, mais dont la résilience est tout de même supérieur ou égale à 50 années.

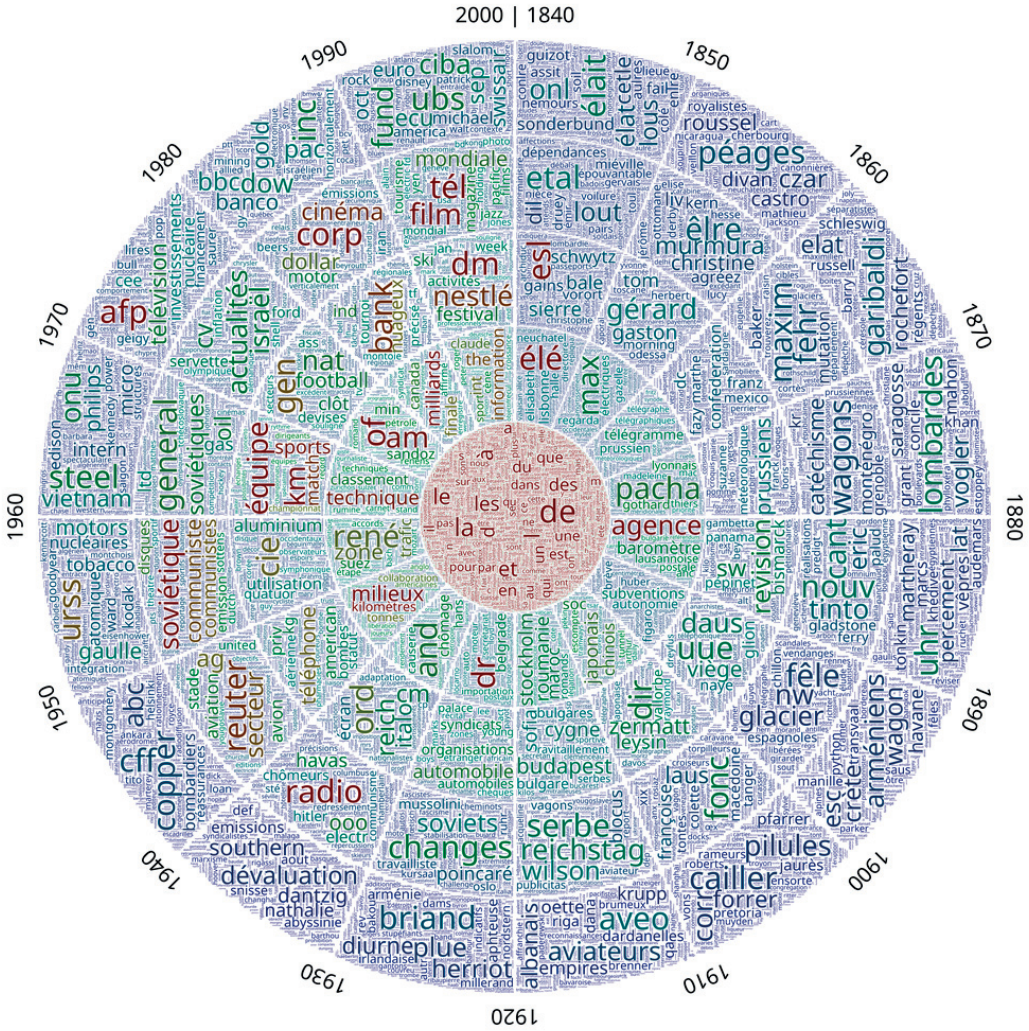


FIGURE 11.37 – Chronocloud de 1-grammes du corpus de GDL

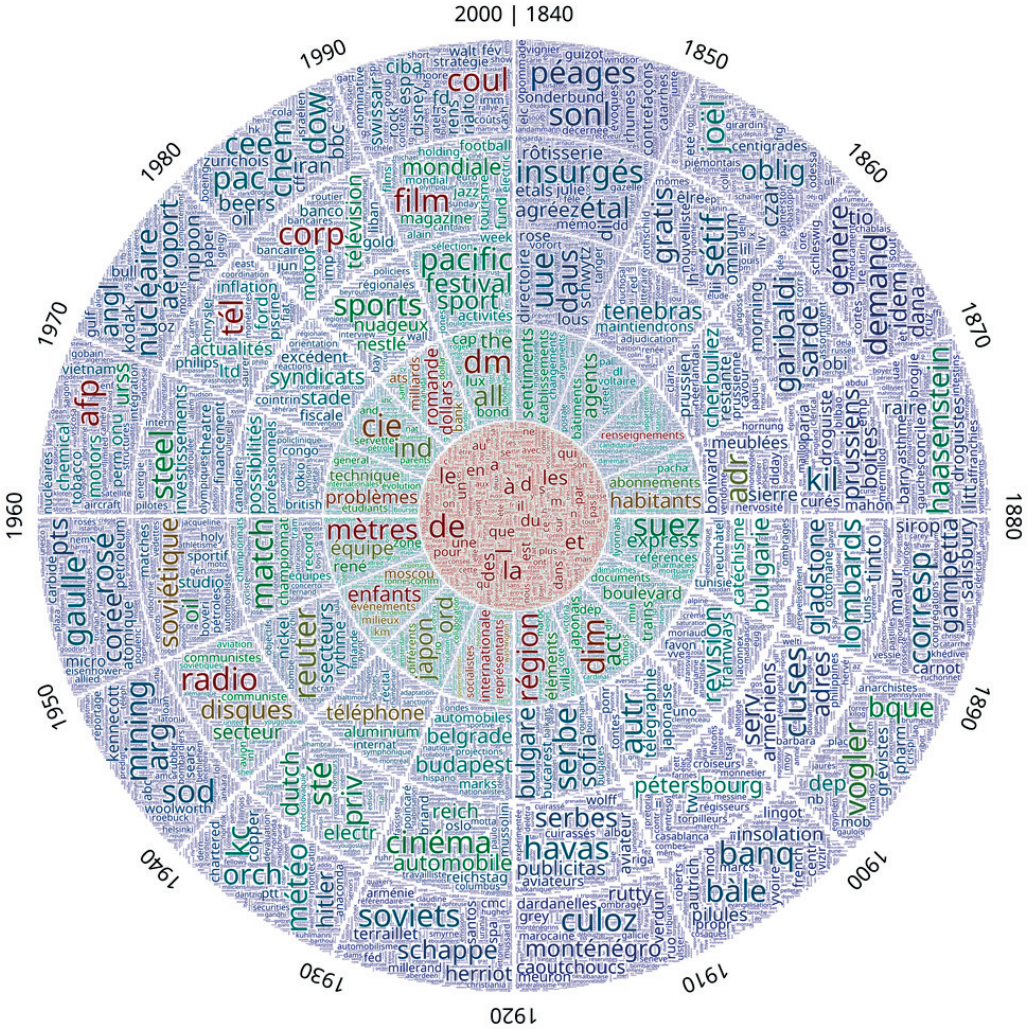


FIGURE 11.38 – Chronocloud de 1-grammes du corpus de JDG

Il n'est pas étonnant de constater visuellement dans les Figures 11.37 et 11.38 que les mots les plus fréquents de l'ensemble de résilience 150 sont similaires dans les deux corpus. On y trouve notamment les mots "de", "l", "à", "la", "le", "les", "d", "des", "et", "a", "que", "qui", "il", soit principalement des articles, conjonctions, prépositions ou pronoms. Il n'est pas anormal de retrouver ces mots au centre du chronocloud, car ils sont usuellement indépendant de contextes particuliers et sont donc spécifiquement liés à la langue étudiée.

En zoomant sur cet ensemble, nous visualisons des mots appartenant au registre journalistique comme ceux relatifs à la politique en général ("gouvernement", "président", "commission", "justice", "assemblée", "constitution", "peuple", "chambre", "ministre" ou simplement

"politique") et aussi à la localisation du journal, sur le plan national ("Suisse", "canton", "confédération", "Conseil", "fédéral" ou "Europe"), mais également de façon plus locale au niveau des cantons ("Lausanne", "Genève", "Fribourg", "vaudois", "Yverdon", "Vevey", "Vaud" ou "romande").

Hors du noyau résilient, nous observons également des mots liés à la politique et à la géographie nationale et internationale, mais aussi aux inventions (notamment audiovisuelles et concernant le transport), à l'économie, aux finances et aux entreprises. Nous observons aussi quelques erreurs d'OCR dans les années plus anciennes (principalement la confusion du "t" avec "l").

En analysant plus précisément les mots appartenant aux couches externes de résilience ($50 \leqslant R < 150$) nous regroupons certains exemples selon des catégories intuitives et tentons de déterminer les raisons probables de leurs positions dans le chronocloud :

– Lieux géographiques : "Bulgarie", "Budapest", "Sofia", "Roumanie", "Maroc", "Belgrade", "URSS", "Israël", "Moscou", "Japon", "Stockholm". Ces exemples de lieux géographiques ne font pas partie du noyau résilient. Ils sont aussi utilisé en relation à des événements historiques et politiques, les plaçant dans des catégories temporelles qui s'expliquent potentiellement au regard de l'histoire. Par exemple, "Stockholm" est placé entre 1910 et 1920 notamment à cause de la conférence de Stockholm de 1917, la troisième et dernière des conférences socialistes contre la guerre. Il est à noter, comme nous le constatons dans la Figure 11.39, que les mots se comportent de façon similaire dans les deux corpus, si ce n'est que les données relatives au corpus de JDG ne contiennent pas les année 1917, 1918 et 1919 rendant le pic de la conférence de Stockholm inexistant dans JDG.

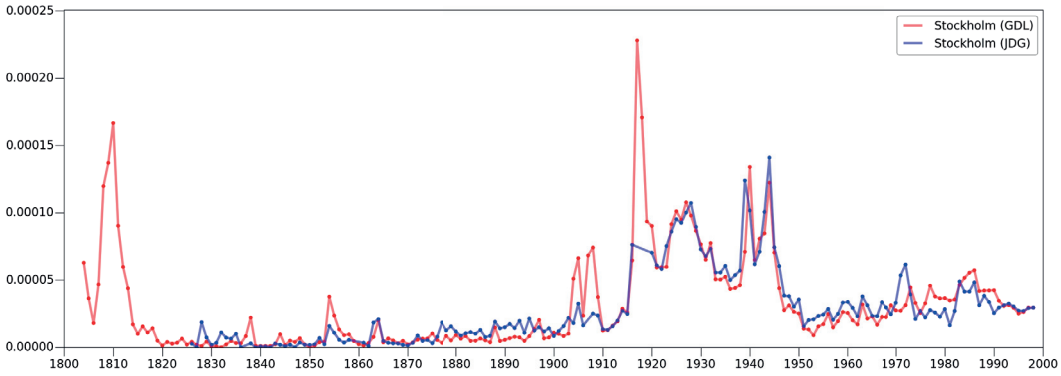


FIGURE 11.39 – Profils fréquentiels du mot "Stockholm"

– Termes référant à des ethnies, nations, etc : "soviétiques", "arméniens", "albanais", "hongrois", "serbes", "canadiens", "chinois", "japonais", "prussiens". Nous remarquons que ces mots sont souvent utilisés pour décrire les actions de troupes dans un contexte de guerre. Par exemple, la position du mot "prussiens" se réfère directement à la guerre franco-allemande (ou guerre franco-prussienne) de 1870 (cf. Figure 11.40).

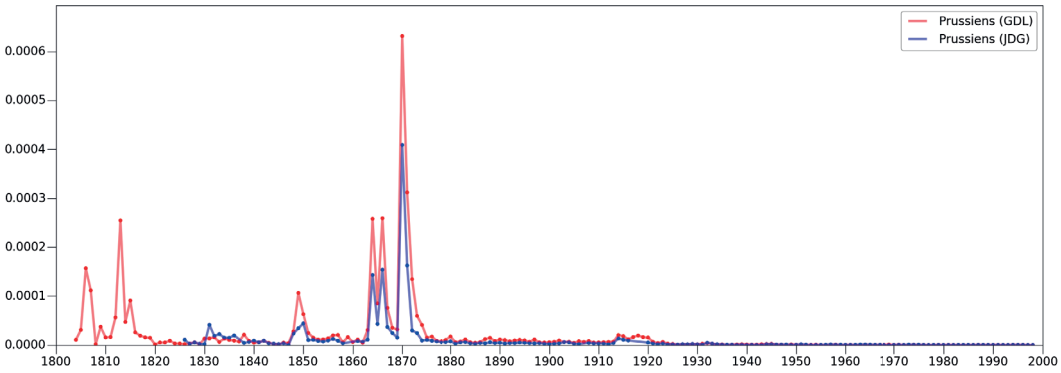


FIGURE 11.40 – Profils fréquentiels du mot "prussiens"

Un autre exemple est donné par les mots "italo" et "abyssin", placé ensemble en 1935 à cause du conflit "italo-abyssin" (appelée aussi italo-éthiopien), entre l'Italie fasciste de Mussolini et l'Empire d'Ethiopie (cf. Figure 11.41).

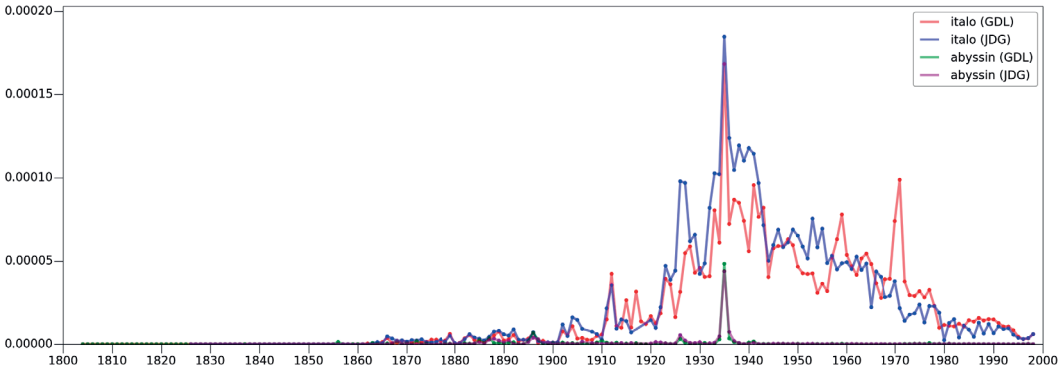


FIGURE 11.41 – Profils fréquentiels des mots "Italo" et "Abyssin"

– Guerres mondiales : "nazisme", "bombes", "bombardier", "collaboration", "Reich", "Reichstag", "Gestapo". Les guerres mondiales ont impacté les deux corpus de journaux. Les pics de fréquence de ces mots se révèlent dans les années de guerre (cf. Figure 11.42).

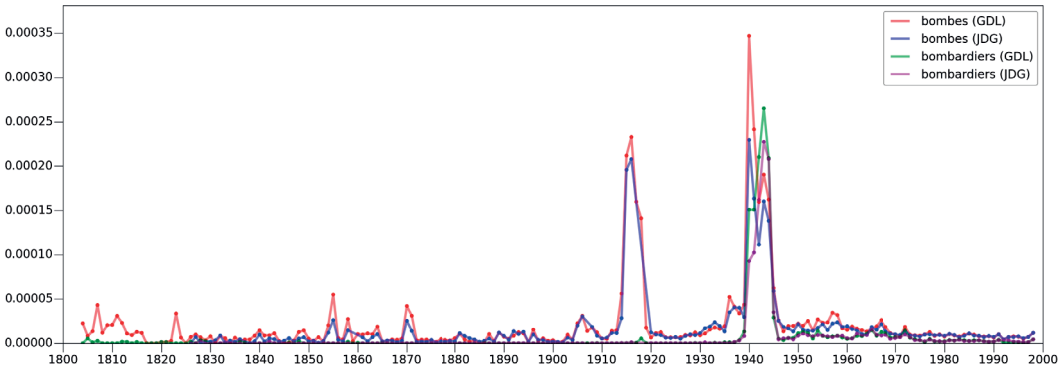


FIGURE 11.42 – Profils fréquentiels des mots "bombes" et "bombardiers"

– Personnalités politiques : "Guizot" (François Guizot), "Mahon" (Patrice de Mac Mahon), "Gaulle" (Charles de Gaulle ou Général de Gaulle), "Druey" (Daniel-Henri Druey), "Herriot" ("Edouard Marie Herriot"), "Briand" (Aristide Briand), "Gladstone" (William Ewart Gladstone), "Garibaldi" (Giuseppe Garibaldi ou Général Garibaldi), "Nemours" (Duc de Nemours), "Mussolini" (Benito Mussolini), "Hitler" (Adolf Hitler), "Castro" (Fidel Castro). Certains de ces noms de personnalités politiques apparaissent sporadiquement, mais en étant résilients sur plus de 50 années (cf. Figure 11.43).

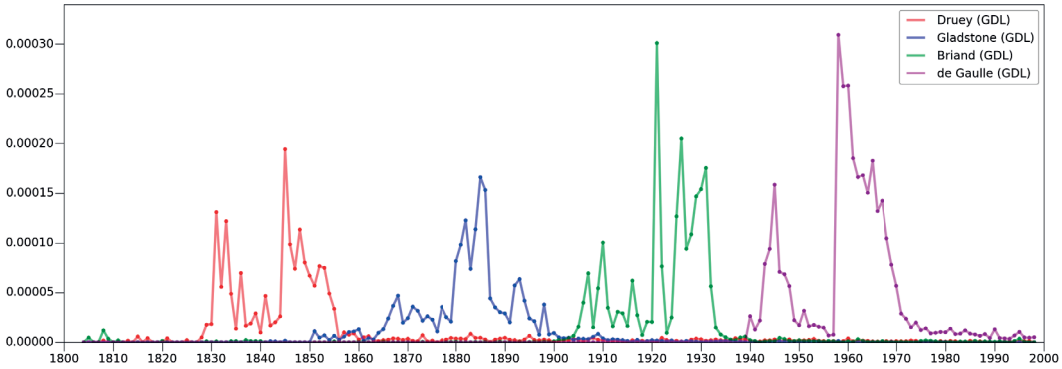


FIGURE 11.43 – Profils fréquentiels des mots "Druey", "Gladstone", "Briand" et "De Gaulle" dans le corpus de GDL

– Entités politiques : "Sonderbund", "Vorort", "ONU", "URSS". Le Sonderbund était une alliance regroupant sept cantons catholiques conservateurs (Fribourg, Lucerne, Schwytz, Unterwald, Uri, Valais et Zoug) afin de se défendre contre la centralisation du pouvoir vers la Confédération suisse. Son profil fréquentiel permet d'observer un pic d'envergure en 1847 correspondant à l'année au cours de laquelle cette alliance fut vaincue. Le Vorort (ou canton directeur) est le directoire fédéral qui était chargé de gérer les affaires en l'absence de l'assemblée des députés des cantons suisses. Son profil fréquentiel permet d'observer la fin de l'utilisation de ce mot d'origine allemande dans le corpus en 1848, année ou ce système prit fin (cf. Figure 11.44).

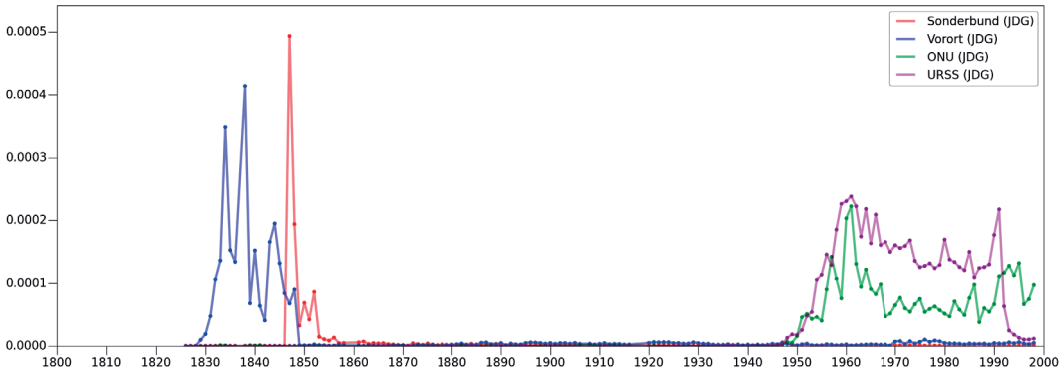


FIGURE 11.44 – Profils fréquentiels des mots "Sonderbund", "Vorort", "ONU" et "URSS" dans le corpus de JDG

– Audiovisuel et communication : "télégramme", "téléphone", "cinéma", "film", "radio", "télévision". La création des technologies audiovisuelles fait également partie de l'histoire du corpus et nous observons à travers la visualisation certaines de ces inventions les plus importantes. Le mot "télégramme" apparaît dans le corpus en 1858, a une fréquence maximale en 1869, puis diminue lentement jusqu'à disparaître. Le mot "télévision" apparaît en 1924, mais une importante augmentation de la fréquence a lieu en 1950. La fréquence maximale est atteinte pendant les années 1964 à 1989 dans les deux corpus. Cette fréquence élevée qui se maintient dans les dernières années du corpus est principalement due au programmes de télévision intégrés dans une section dédiée du journal (cf. Figure 11.45).

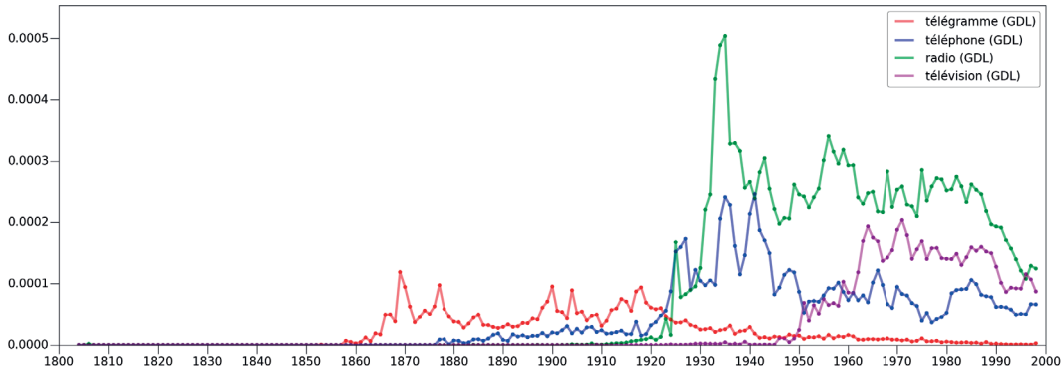


FIGURE 11.45 – Profils fréquentiels des mots "télégramme", "téléphone", "radio" et "télévision" dans le corpus de GDL

– Transport : "tram", "tramway", "automobile", "aviation", "avion", "aéroport". Les moyens de transport ont également une chronologie claire dans les deux corpus (cf. Figure 11.46).

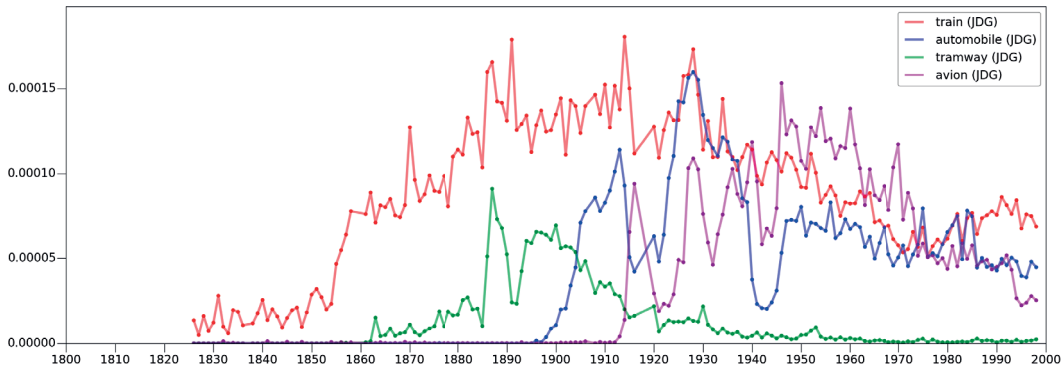


FIGURE 11.46 – Profils fréquentiels des mots "télégramme", "téléphone", "radio" et "télévision" dans le corpus de JDG

– Sports : "sports", "équipe", "olympique", "football", "Servette", "tennis", "hockey". Ces mots sont principalement placés dans le dernier quadrant (années 1960 à 2000), résultant de l'augmentation de la place du sport dans les journaux à mettre en parallèle avec un contexte international d'importance croissante (cf. Figure 11.47).

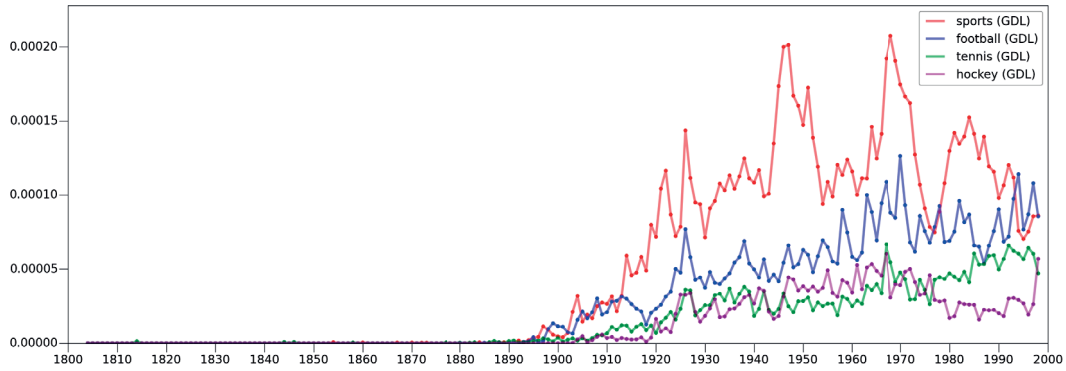


FIGURE 11.47 – Profils fréquentiels des mots "sports", "football", "tennis" et "hockey" dans le corpus de GDL

– Sociétés et bourse : "Reuters", "Nestlé", "Cailler", "Motors" (General Motors), "Ford", "Kodak", "Ciba" (Ciba-Geigy / Ciba Specialty Chemicals), "UBS", "Disney". Ces mots apparaissent également en majorité sur le dernier quadrant du chronocloud de 1960 à 2000. Leurs importance est due au développement du secteur de l'industrie et la tendance globale à la mondialisation. De plus, une section particulière du journal consacrée à la bourse aura tendance à reprendre systématiquement ces termes augmentant encore leur importance dans les années les plus récentes. La marque de chocolat Cailler est ancienne, fondée en 1819, et apparaît donc très tôt dans les pages de la bourse (cf. Figure 11.48).

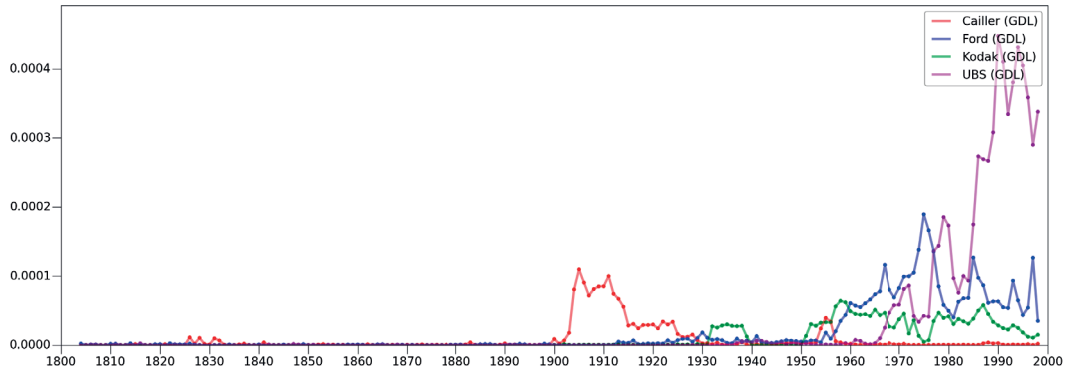


FIGURE 11.48 – Profils fréquentiels des mots "cailler", "ford", "kodak" et "ubs" dans le corpus de GDL

– Erreurs d'OCR récurrentes : "élé", "esl", "lout", "étal","élait", "onl". Tous ces mots, plus visibles dans le chronocloud de GDL plutôt que celui de JDG, sont proches de mots français communs via le remplacement de la lettre "l" par "t" suggérant que le système OCR utilisé a rencontré des difficultés à transcrire ces mots dans les années les plus anciennes en raison de la confusion probable entre les deux lettres. En effet, les changements de police opérés par le journal selon différentes périodes ont pu induire cette erreur, en particulier en cas d'écriture plus petite et de police différente. Leurs fréquences maximales les placent majoritairement entre les années 1840 et 1850 dans le chronocloud.

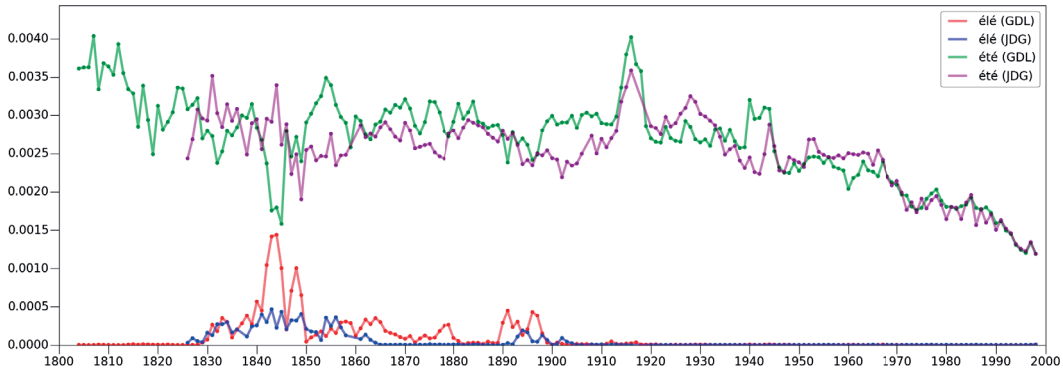


FIGURE 11.49 – Profils fréquentiels des mots "élé" et "été"

Nous observons dans la Figure 11.49, les profils fréquentiels de "élé" et son correspondant supposé, "été". De façon intéressante, nous constatons que le corpus de GDL est effectivement plus touché par ce type d'erreur d'OCR dans une période allant de 1842 à 1849 avec une exception lors de l'année 1846. Nous observons également que l'augmentation de fréquence du mot "élé" semble corrélé avec une diminution de la fréquence du mot "été" renforçant l'hypothèse d'un taux d'erreur plus élevé de transcription de la lettre "t" par la lettre "l" dans cette période particulière.

– Diverses abréviations : "kil" (pour kilos), "dr" (pour docteur), "cie" (pour compagnie), "tél" (pour téléphonne), "kc" (pour la fréquence d'onde radio), "dim" pour dimanche, "orch" pour orchestre. Ces abréviations sont rencontrées dans des sections particulières du journal comme les annonces, la bourse, les événements culturels, les horaires de cinéma ou les chaînes de radio (cf. Figure 11.50).

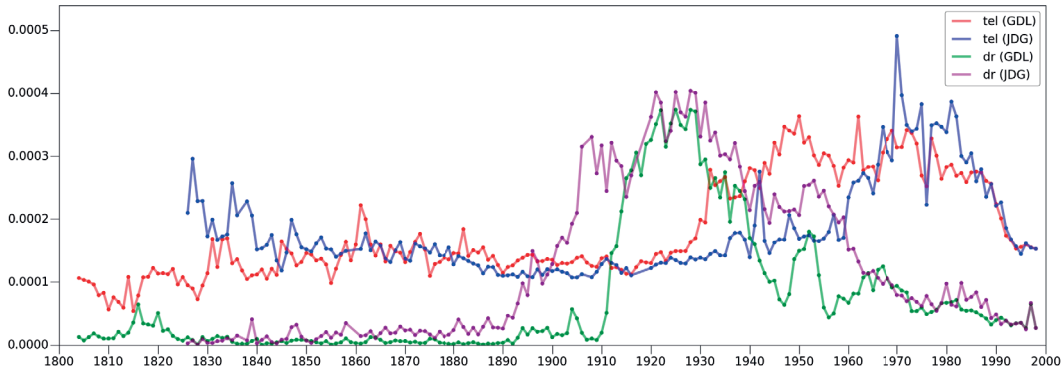


FIGURE 11.50 – Profils fréquentiels des mots "tél" et "dr"

Beaucoup d'autres mots particuliers peuvent être trouvés notamment en utilisant l'option de zoom de l'interface en ligne. C'est mots se révèlent souvent intéressants et leur position peut généralement être expliquée par recherche directe dans les archives. Toutefois, ces recherches ne constituent pas des preuves irréfutables, mais plutôt des indices et éléments qui mis ensemble permettent de corroborer ou non une hypothèse.

Un problème général que peut poser ce type d'analyse est lié à la polysémie, soit l'ambiguïté du sens rattaché à un mot. Par exemple, le mot "Castro" est placé dans deux différentes catégories temporelles selon qu'on analyse le corpus JDG ou GDL et ces catégories sont éloignées. Dans les deux corpus, il y a une fréquence élevé du mot en 1960 lorsque Fidel Castro se rapproche de l'Union des Républiques Socialistes Soviétiques. Cependant, dans le corpus de GDL uniquement, il existe un autre pic de fréquence plus élevé. Une recherche dans les archives montre qu'il s'agit de "madame de Castro" ou du "comte de Castro" qui fait partie d'une histoire écrite dans une section spéciale du journal (chaque parution révèle une partie de l'histoire commencée lors d'une précédente parution). La notion de polysémie rend certains profils fréquentiels décomposables selon le nombre de sens que l'on peut y rattacher. Cela peut donc poser problème à un niveau $n = 1$. Toutefois, le problème de la polysémie aura tendance à se résoudre à des niveaux supérieurs $n > 1$, car une combinaison de mots permet en général de préciser le sens que l'on désire y rattacher.

Nous observons également un fait intéressant concernant la tendance générale des mots de résilience $50 \leqslant R < 150$ autant dans les Figures 11.37 et 11.38 représentant les corpus de GDL et JDG que dans les Figures 9.16 à 9.24 représentant les corpus de Google Books pour toutes les langues : dans la catégorie des mots de résilience $50 \leqslant R < 150$, les mots sont d'autant moins fréquents en moyenne qu'il sont anciens. Cela suggère que les mots apparaissant dans les années récentes et se maintenant au moins 50 années successives dans le corpus auront tendance à avoir des pics de fréquence plus élevés. Cette caractéristique s'observe de façon immédiate au niveau de la couleur des mots qui lie les fréquences sur l'entier du chronocloud, passant d'une tendance de couleur froide (faible fréquence) dans les années anciennes jusqu'aux couleurs plus chaudes (hautes fréquence) dans les années récentes. Nous représentons le nombre de mots de résilience $50 \leqslant R < 150$ (gauche) et la somme de leurs fréquences pour les 100 mots les plus fréquents, par année de fréquence maximale pour les corpus de GDL et JDG dans la Figure 11.51. Nous y constatons un nombre élevé de mots pour les années de fréquence maximale anciennes et récentes. Toutefois, la fréquence des mots les plus fréquents est généralement plus élevée dans les années récentes.

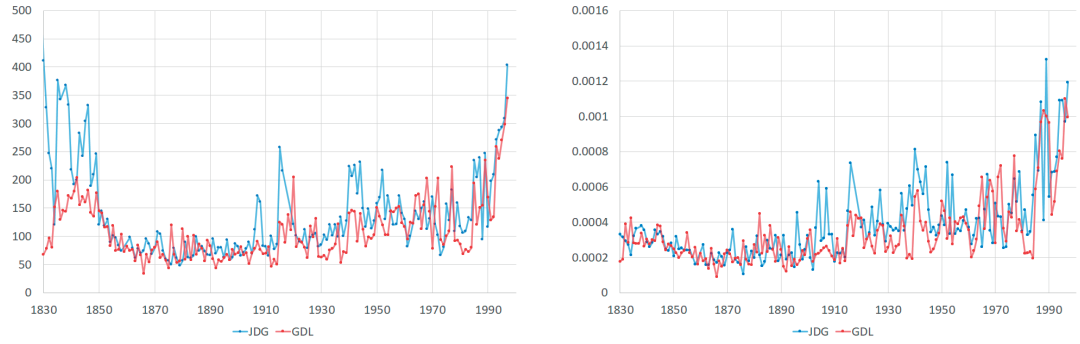


FIGURE 11.51 – Nombre de mots (gauche) et la somme des fréquences des 100 mots les plus fréquents, de résilience $50 \leqslant R < 150$, par année de fréquence maximale pour les corpus de GDL et JDG

Chronocloud différentiels

Dans cette section, nous utilisons la notion de chronoclouds différentiels afin d'observer les différences majeures entre les deux corpus en termes de profils fréquentiels des mots composant ces corpus. Afin de visualiser directement le sens de la variation de fréquences, nous utilisons la variante asymétrique.

Le chronocloud différentiel asymétrique de GDL moins JDG est représenté dans la Figure 11.52 et celui de JDG moins GDL est représenté dans la Figure 11.53.

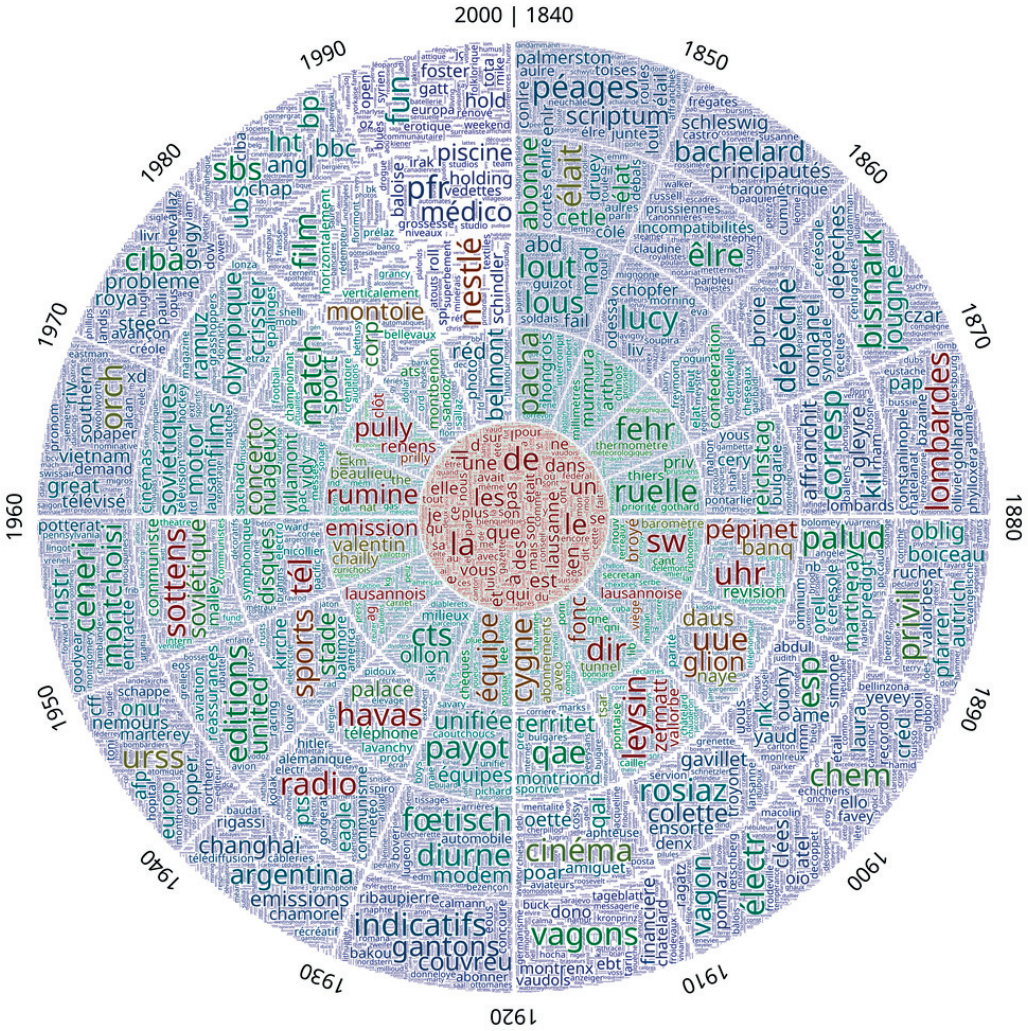


FIGURE 11.52 – Chronocloud différentiel asymétrique de 1-grammes du corpus de GDL moins celui de JDG

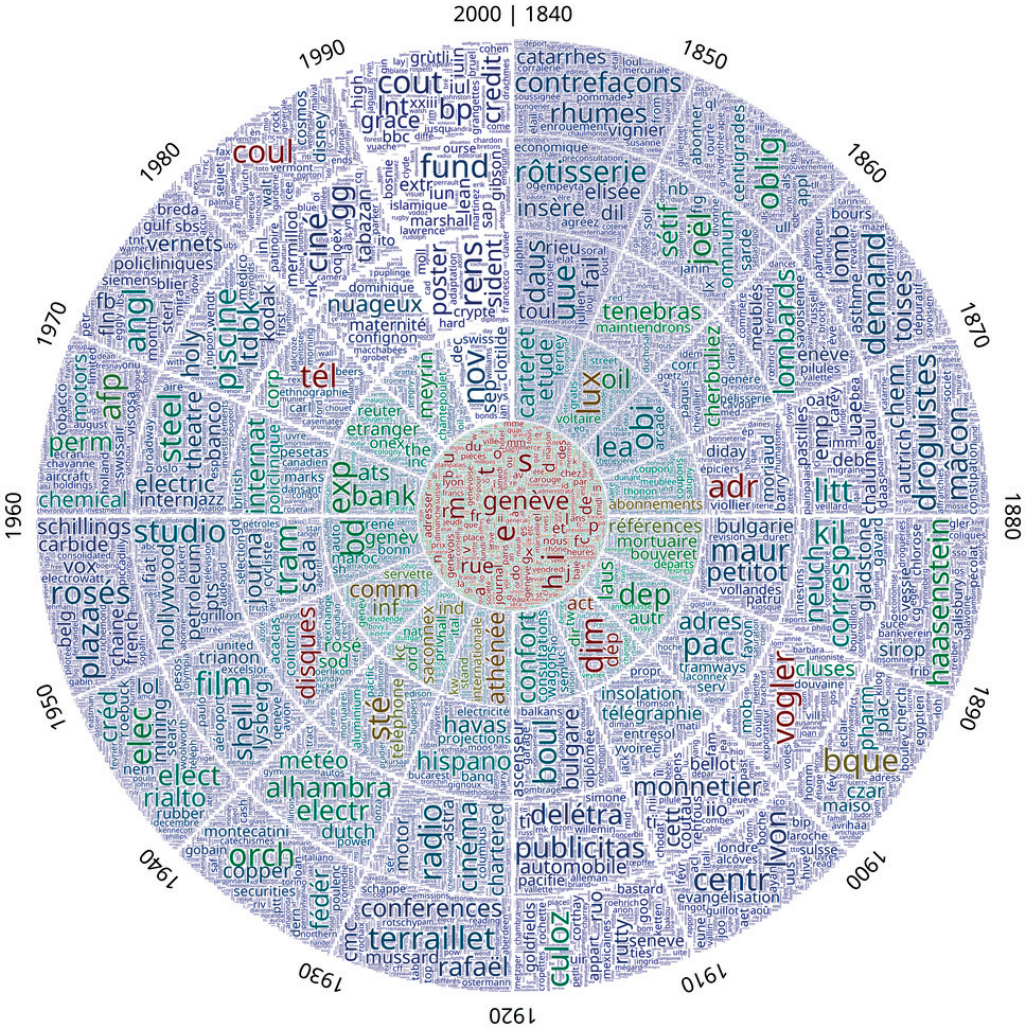


FIGURE 11.53 – Chronocloud différentiel asymétrique de 1-grammes du corpus de JDG moins celui de GDL

La plupart des mots visibles dans ces représentations sont des lieux géographiques, des adresses, des corporations ou des personnes connues, mais à un niveau local. En effet, les deux journaux ont une différence de localisation (Genève / Lausanne) qui induit des différences évidentes de profil fréquentiel au niveau des lieux et personnes rattachés à la région.

Ces chronoclouds font également ressortir d'autres types d'erreurs d'OCR notamment celles résultant de la confusion probable des lettres "u" et "n" ou bien "y" et "v". Ces différences résultent aussi du fait que la police utilisée dans un journal à un moment donné est souvent différente que celle utilisée par l'autre journal au même moment.

Nous observons également que des mots correspondant à des métiers, comme "monteur" ou "tanneur", sont inégalement fréquents dans les deux journaux, mais globalement en raison de publicités et annonces incorporées dans le corpus plutôt que de différences purement linguistiques (cf. Figure 11.54).

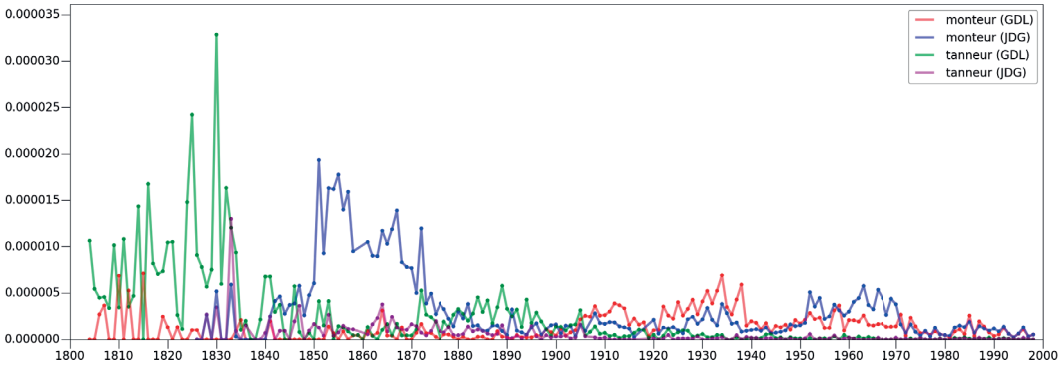


FIGURE 11.54 – Profils fréquentiels des mots "monteur" et "tanneur"

Le mot "tanneur" fait référence au métier permettant le travail de la peau d'un animal à l'aide de tanin pour la transformer en cuir. Ce mot n'est repris que dans le corpus de GDL.

Le mot "monteur" fait principalement référence au métier de "monteur de boîtes", toutefois la précision "de boîtes" est omise dans les années plus récentes, comme le montre la comparaison des profils fréquentiels du mot "monteur" et du 3-grammes "monteur de boîtes" qui correspondent presque exactement pour la première moitié du corpus (cf. Figure 11.55).

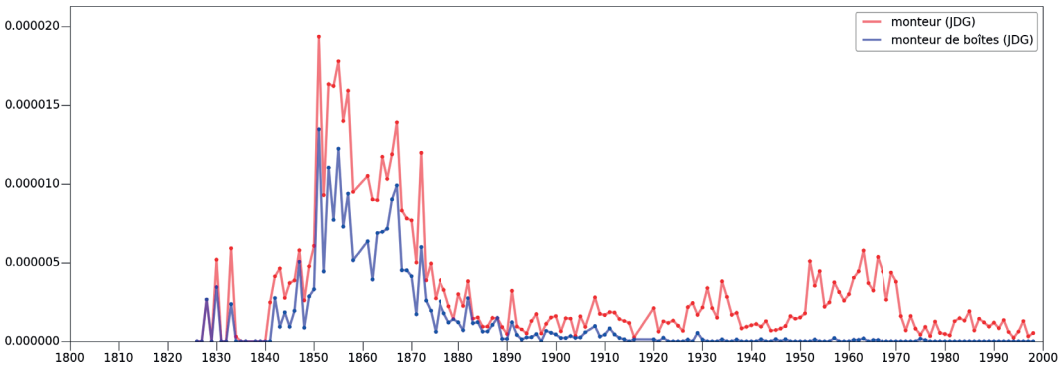


FIGURE 11.55 – Profils fréquentiels du mot "monteur" et du 3-gramme "monteur de boîtes" pour JDG

La relation qu'entretiennent ces deux profils fréquentiels est intéressante, car elle montre la correspondance entre deux profils fréquentiels à une période donnée et leur divergence dans les années les plus récentes.

Il est intéressant de constater que ce recouvrement fréquentiel entre "monteur" et "monteur de boîtes" indique que le mot "monteur" n'est quasiment jamais utilisé autrement qu'avec la suite de mots "de boîtes". Cet effet s'estompant avec les années, le 3-gramme "monteur de boîtes" tombe à une fréquence quasi nulle tandis que monteur continue tout de même d'être utilisé dans les années plus récentes.

Certaines sections comme les "Avis officiels, juridiques et administratifs" (présents dans le corpus de GDL) utilisent des mots comme "toise", c'est notamment le cas du journal GDL dans les années antérieures à 1835. La toise est une ancienne unité de mesure qui correspond à 6 pieds, soit à peu près la longueur de deux bras tendus en prenant la mesure depuis le bout des doigts (cf. Figure 11.56).

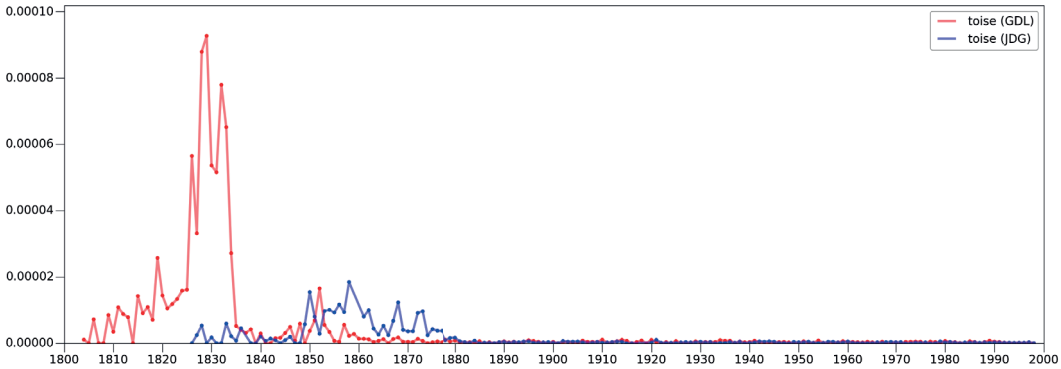


FIGURE 11.56 – Profils fréquentiels du mot "toise"

Une différence typique due à la localisation des deux journaux est l'évocation de la population locale "lausannois" et "genevois" en terme de fréquence.

Le journal GDL évoque les "genevois" presqu'aussi souvent que les "lausannois" ce qui est loin d'être le cas du journal JDG (cf. Figure 11.57).

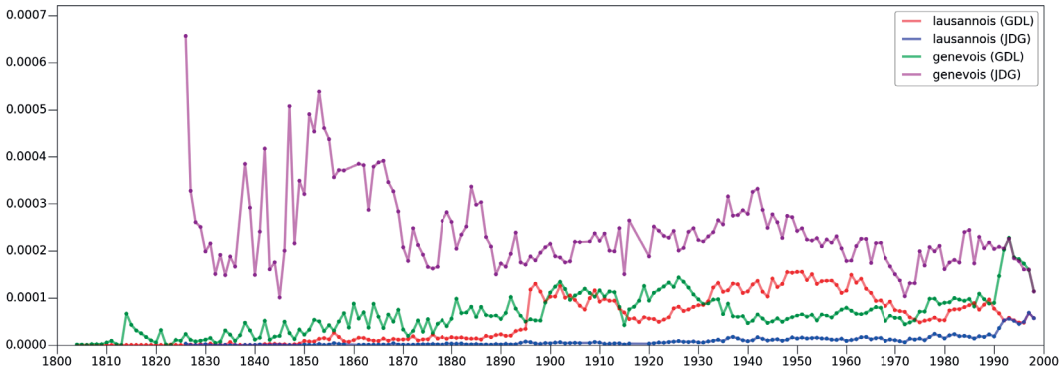


FIGURE 11.57 – Profils fréquentiels des mots "lausannois" et "genevois"

Outre ces différences causées par une sur-représentation de certains mots dans diverses sections du journal et les mots considérés comme "externe à la langue" comme les lieux, noms de personne, compagnies et corporations, le chronocloud différentiel montre quelques exemples d'usages linguistiques différents entre les deux journaux. Ces effets découlent notamment de décisions éditoriales différentes entre les deux journaux en termes de références orthographiques pour certains mots.

Par exemple, "vagon" était autrefois une écriture alternative de "wagon". La différence de profil fréquentiel montrée par le chronocloud différentiel s'explique par le choix du journal GDL de commencer à utiliser l'orthographe "vagon" au lieu de "wagon" dès 1899 et de l'adopter pleinement en 1904. Cependant, l'autre journal n'a pas adopté ce choix linguistique et la forme "vagon" a donc gardé une faible fréquence pendant toute la période couverte par le corpus de JDG. Dans le cas de GDL, la forme "vagon" a simplement remplacé "wagon" et ce remplacement s'est terminé 70 ans plus tard quand l'utilisation de "vagon" est devenu obsolète. Les profils fréquentiels de "wagon" et "vagon" pour les corpus de JDG et GDL sont présentés dans les Figures 11.58 et 11.59.

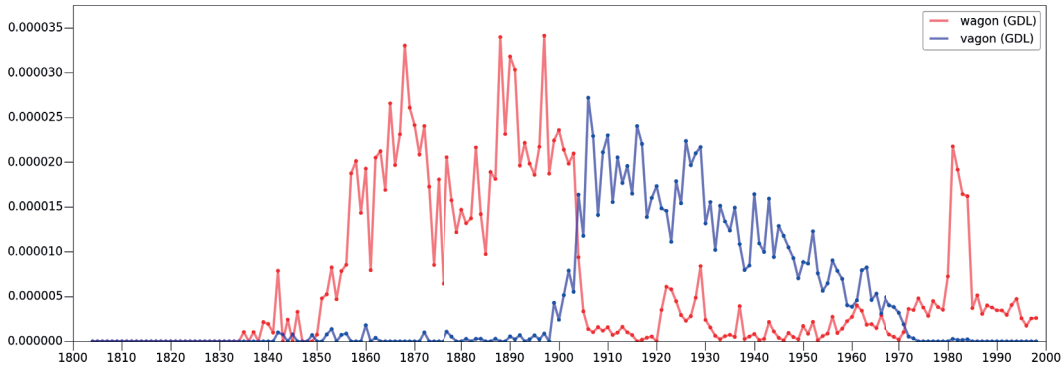


FIGURE 11.58 – Profils fréquentiels de "wagon" et "vagon" pour GDL

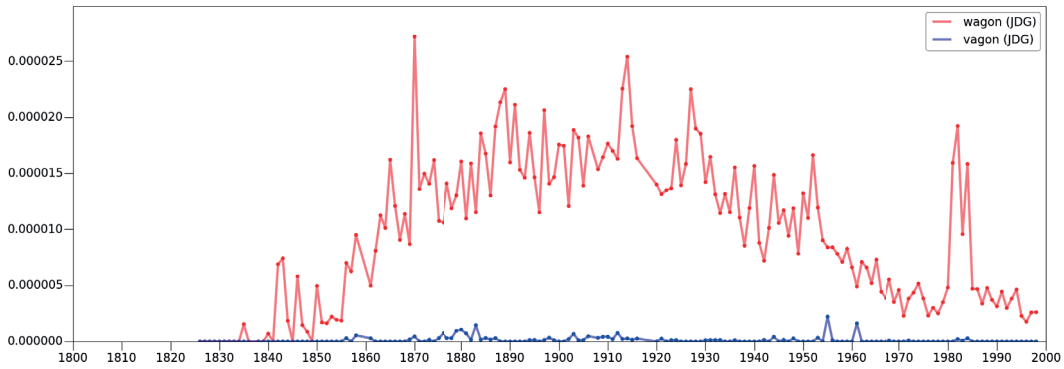


FIGURE 11.59 – Profils fréquentiels de "wagon" et "vagon" pour JDG

Un autre exemple de ce même type peut être repéré sur le chronocloud différentiel avec la forme "Changaï" utilisée par le journal GDL au lieu de "Shangaï". Encore une fois, il apparaît un remplacement assez évident dans le journal GDL uniquement, désignant la ville de Shanghaï par la forme "Changhaï" entre 1925 et 1985.

Le mot est toutefois trop peu présent dans les années plus récentes des deux corpus pour pouvoir observer le moment où la forme "Shanghaï" aurait de nouveau pris le dessus sur celle de "Changhaï" sachant que cette forme est devenu désuète aujourd'hui et que seule la forme "Shanghaï" subsiste.

Pourtant, il est intéressant de constater que la forme "Changhaï" n'a jamais été reprise par JDG. Le remplacement entre les deux versions "Shangaï" et "Changaï" pour GDL peut être observé dans la Figure 11.60 à comparer avec la Figure 11.61 représentant les profils fréquentiels des même formes, mais dans le corpus de JDG.

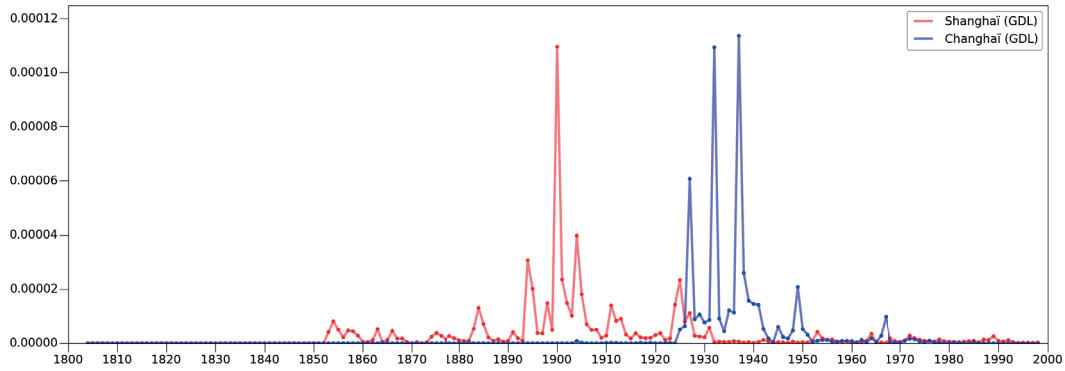


FIGURE 11.60 – Profils fréquentiels de "Shanghaï" et "Changhaï" pour GDL

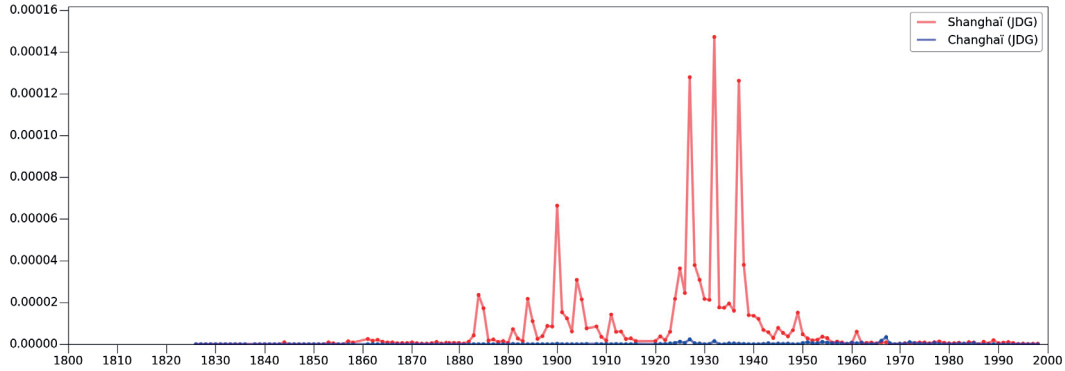


FIGURE 11.61 – Profils fréquentiels de "Shanghaï" et "Changhaï" pour JDG

Nous observons bien un remplacement clair pour le corpus de GDL tandis que le journal JDG opte pour garder la première forme, celle de "Changhaï" étant inexistante dans JDG.

Un autre exemple appartenant au même registre est celui de la ville de Tokyo. C'est la forme "Tokio" qui était utilisée dans les corpus de JDG et GDL dans les années antérieures 1965 alors que la forme "Tokyo" apparait vers 1960 et fini par remplacer la première forme.

Les profils fréquentiels de "Tokio" et "Tokyo" pour les corpus de JDG et GDL sont présentés dans les Figures 11.62 et 11.63.

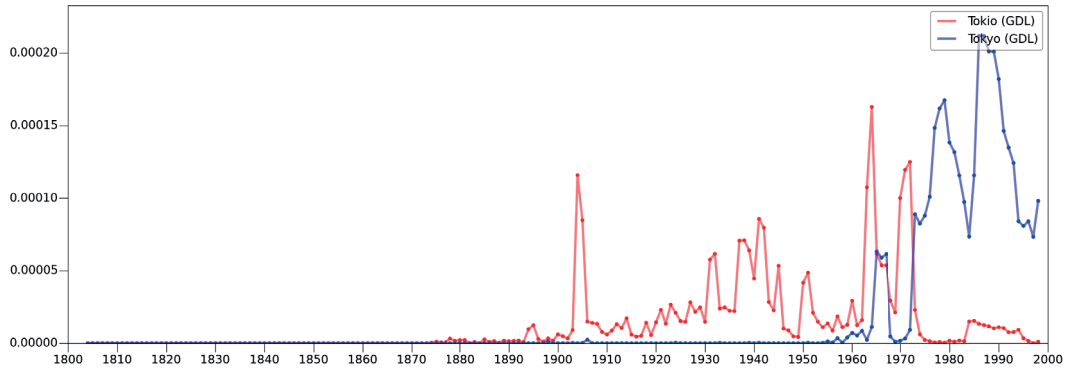


FIGURE 11.62 – Profils fréquentiels de "Tokio" et "Tokyo" pour GDL

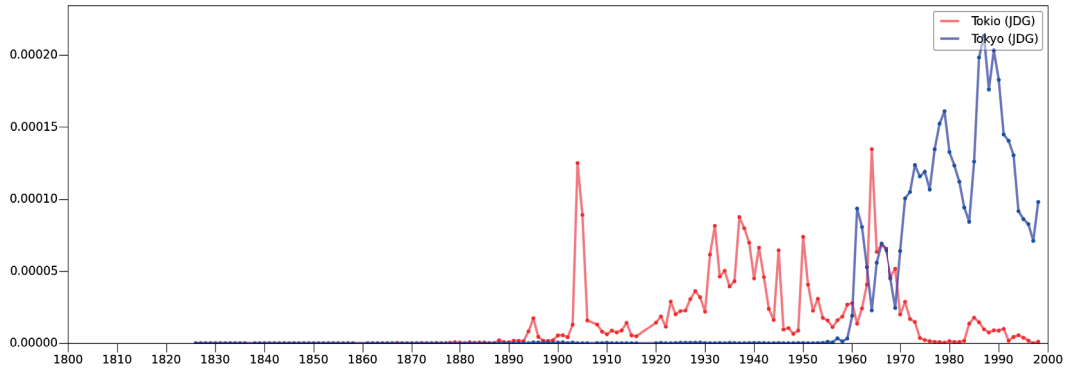


FIGURE 11.63 – Profils fréquentiels de "Tokio" et "Tokyo" pour JDG

Nous observons que ce remplacement a lieu dans les deux corpus dans la même période. Le phénomène est progressif et sa durée est, dans ce cas de figure, d'une quinzaine d'années avant qu'une forme ait totalement remplacé l'autre.

Un dernier exemple de ce type, mis en évidence par le chronocloud différentiel, est l'utilisation des formes "Tsar" et "Czar" désignant anciennement le souverain de la Russie. "Czar" est l'ancienne version, aujourd'hui désuète, de "Tsar". Les profils fréquentiels de ces formes sont représentés dans les Figures 11.64 et 11.65.

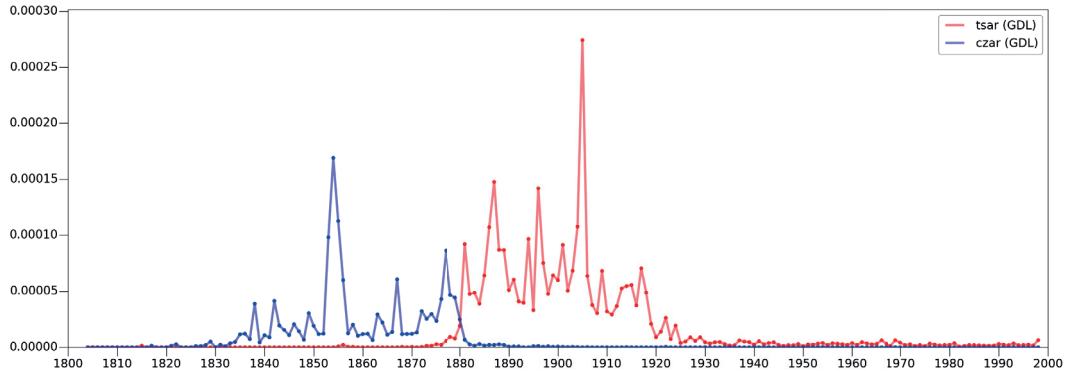


FIGURE 11.64 – Profils fréquentiels des mots "Tsar" et "Czar" pour GDL

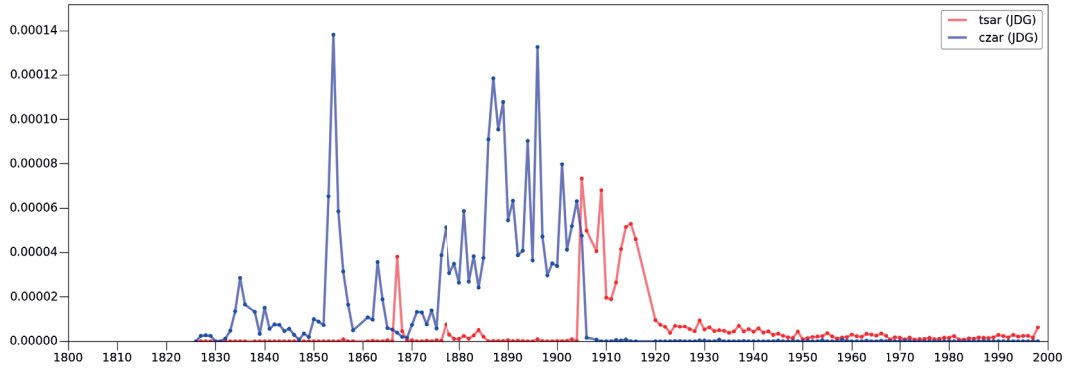


FIGURE 11.65 – Profils fréquentiels des mots "Tsar" et "Czar" pour JDG

Dans cet exemple, les deux journaux ont décidé de changer d'orthographe. Toutefois, il ne l'ont pas fait en même temps et GDL en a été le précurseur, décidant le changement en 1880 tandis que JDG l'a fait en 1905. Ces différents exemples de remplacement dans un corpus donné d'une forme par une autre nous permettent d'observer l'effet de la concurrence entre deux formes pour une même signification sur des corpus journalistiques. Il est clair que ces effets sont minimes dans une vision macroscopique du langage et les techniques d'étude globale de la langue en considérant des distances entre les sous-corpus pour chaque année ne seront que peu impactées.

Toutefois, le chronocloud est un outil qui fait le lien entre la vision Macro et Micro du corpus et c'est ce qui permet de mettre en évidence ces mots au comportement étranges et intéressants. La comparaison du profil fréquentiel nous permet ensuite d'observer ce phénomène avec précision. Le fait de disposer de deux corpus journalistiques très proches permet en outre d'affiner nos observations, car la somme des profils fréquentiels de deux formes en concurrence est généralement la même dans les deux journaux.

11.4 Visualisateur de n-grammes

L'analyse des chronoclouds permet de découvrir rapidement des mots intéressants dans un large corpus. Ces mots peuvent ensuite être analysés plus précisément en utilisant le visualisateur de profils fréquentiels.

Ce chapitre est centré sur l'analyse "au microscope" de l'évolution des mots particuliers du lexique de GDL et JDG. Certains de ces mots et profils fréquentiels sont repérés intuitivement par le croisement de différentes recherches effectuées directement sur le visualisateur tandis que d'autres sont suggérées par les visualisations chronocloud comme dans la section précédente.

Classiquement, il est possible d'étudier le comportement diachronique des mots dans les deux corpus en naviguant simplement dans l'espace des mots selon leurs significations supposées ou selon une classification préétablie sur la base de nos connaissances. Il s'agit d'ailleurs de l'utilisation la plus intuitive du visualisateur de n-grammes, permettant d'amener des éléments de réponses à diverses questions de recherches, historiques comme sociolinguistiques, concernant les corpus étudiés.

Toutefois, il est possible de chercher aussi dans l'espace des courbes de profils fréquentiels des n-grammes. Afin d'illustrer la différence entre ces deux espaces, nous considérons les 96 mots les plus fréquents de chaque corpus. Nous les classons ensuite par ordre de fréquence décroissante et les présentons dans les Tables 11.1 et 11.2.

Ces mots correspondent en grande partie à ceux observés dans le centre du chronocloud de chaque corpus (Figures 11.37 et 11.38). En effet, même si le chronocloud utilise le critère de la résilience des mots pour déterminer leurs positions, il est clair que les mots ayant une haute fréquence auront tendance à avoir également une grande résilience. Il faut toutefois noter, qu'il s'agit d'une tendance moyenne et non d'une règle absolue.

de	que	m	son	4	leur	0	dont
la	un	on	été	y	deux	h	t
le	il	n	nous	sa	vous	e	30
à	une	ce	aux	ses	si	7	lausanne
l	est	se	5	8	comme	conseil	faire
et	qui	ne	cette	je	fait	bien	20
les	dans	sur	sont	lui	même	j	50
d	pour	pas	elle	i	suisse	tous	9
des	au	plus	c	ou	était	sans	00
du	s	1	3	ces	6	après	encore
a	par	ont	2	être	ils	où	grand
en	qu	avec	mais	tout	avait	10	fr

TABLE 11.1 – 96 mots les plus fréquents du corpus de GDL classés par ordre de fréquence décroissante par colonne

La possibilité d'interpréter les Tables 11.1 et 11.2 repose sur une connaissance préétablie au moins partielle de la langue française. Autrement dit, il est possible de chercher dans cet espace si le chercheur connaît déjà la signification des signes et qu'il possède une question de recherche précise pour ensuite regarder les profils fréquentiels et en tirer éventuellement des conclusions.

Nous observons par exemple qu'il y a des mots, des lettres seules et des nombres. Ces mots sont particulièrement communs au sein de la langue française et appartiennent principalement aux catégories : déterminant, pronom ou préposition (comme constaté antérieurement via la visualisation chronocloud).

Les quatre mots "suisse", "conseil", "lausanne" et "genève" sont assez particuliers. En effet, il est clair que les raisons de leur présence parmi les mots les plus fréquents sont liées à la localisation du journal. Nous observons aussi la présence des verbes "avoir" et "être", notamment conjugués à la troisième personne de l'imparfait.

Toutefois, si nous avions regardé le même tableau dans une langue totalement inconnue, il aurait été hasardeux de tirer des conclusions quant à la nature de ces mots.

Afin de visualiser un autre espace de recherche, celui des courbes de profils fréquentiels, nous transposons les mêmes Tables 11.1 et 11.2 en tables de profils fréquentiels, remplaçant simplement le mot par son profil fréquentiel dans les Figures 11.66 et 11.67.

Nous utilisons la visualisation appelée "small multiple" (Raveneau, 1993) permettant de montrer l'évolution globale de plusieurs courbes en se focalisant sur les tendances plus que sur les valeurs. Les axes et leurs indications sont donc omis afin de laisser la place à l'interprétation des courbes tout en sachant qu'elles sont toutes basées sur le même axe temporel en abscisse et sur un axe fréquentiel variable en ordonnée.

de	que	qu	c	sont	ou	tout	fr
la	un	n	i	cette	ses	25	00
l	il	on	ont	mais	30	fait	sans
le	une	1	3	8	j	deux	15
à	est	ce	avec	elle	genève	lui	avait
et	qui	pas	2	t	leur	9	12
les	s	ne	4	10	si	ils	était
d	dans	se	été	y	ces	conseil	tous
des	pour	sur	e	50	suisse	je	p
a	m	plus	son	7	être	0	dont
du	au	5	aux	6	20	même	r
en	par	nous	h	sa	comme	bien	vous

TABLE 11.2 – 96 mots les plus fréquents du corpus de JDG classés par ordre de fréquence décroissante par colonne

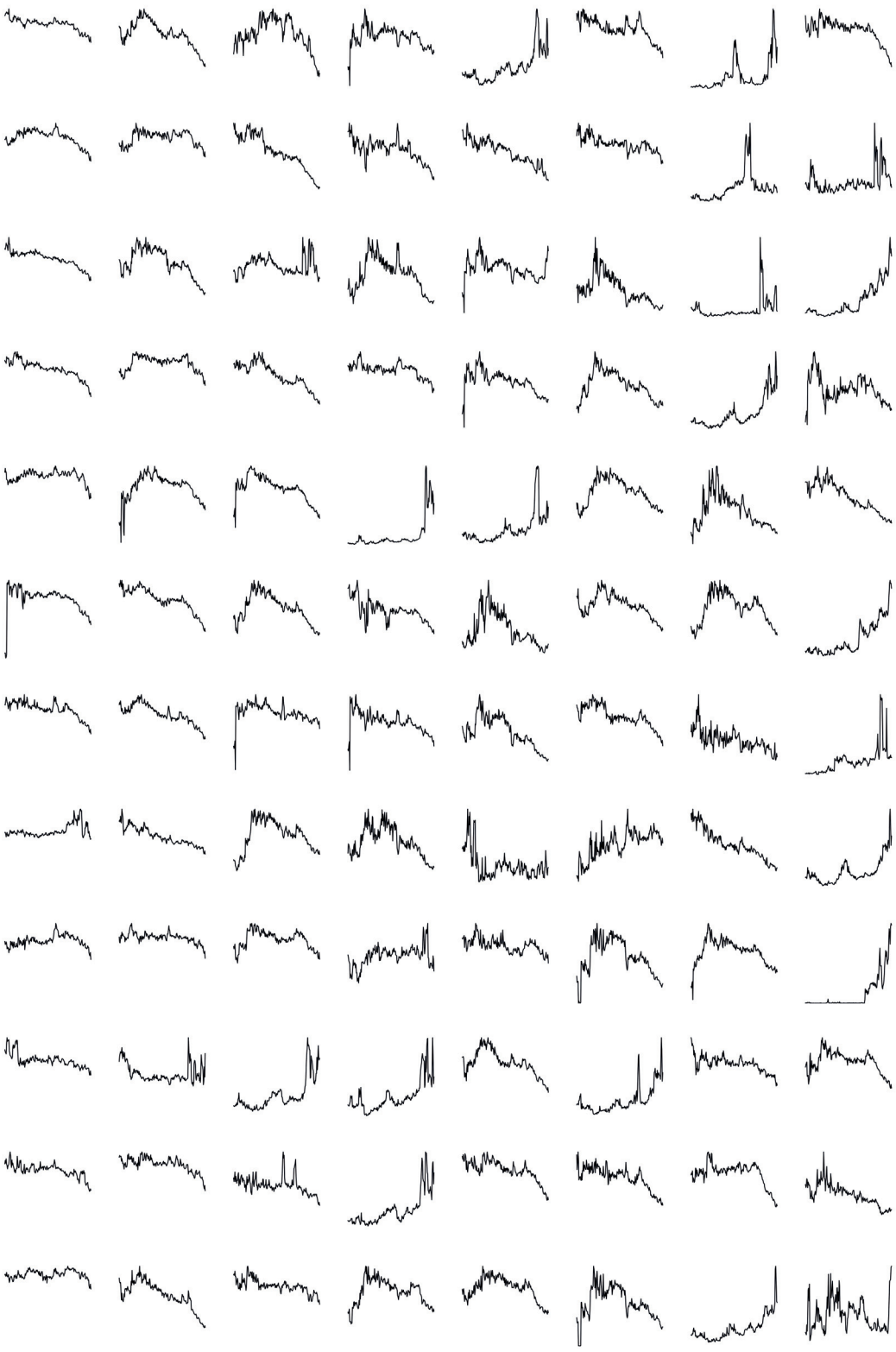


FIGURE 11.66 – 96 profils fréquentiels du corpus de GDL correspondant aux mots de la Table 11.1 dans le même ordre

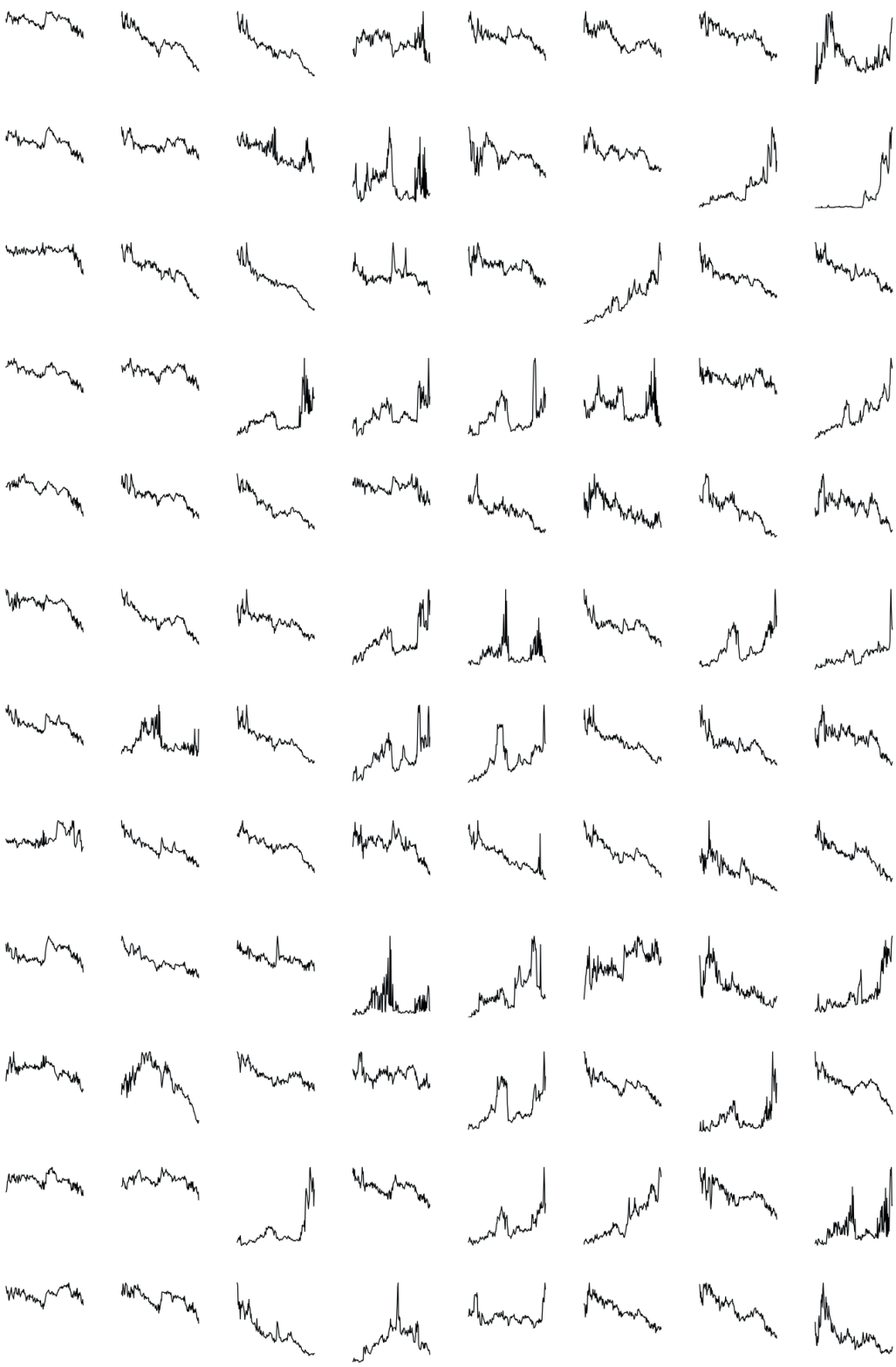


FIGURE 11.67 – 96 profils fréquentiels du corpus de JDG correspondant aux mots de la Table 11.2 dans le même ordre

Nous observons cette fois l'ensemble des entités en connaissant uniquement leurs variations fréquentielles diachroniques et leur ordre fréquentiel. L'interprétation de cet espace est complètement indépendant de la langue. Nous y observons des profils fréquentiels stables, instables, croissants, décroissants, de forme relativement convexe ou même plutôt concave. Cette diversité dans les profils fréquentiels de mots les plus fréquents est la dimentionalité de l'information que peut contenir une telle table, permettant de repérer des mots particuliers sans les connaître.

Il est donc possible de rechercher les mots sur l'espace des courbes de profil fréquentiel plutôt que celui des mots eux-mêmes. Nous pouvons distinguer deux types d'observations concernant la recherche axée sur les profils fréquentiels :

– Caractérisation de la typologie des courbes de profils fréquentiels et classification des mots en fonction de leur catégorie.

– Relations particulières entre deux ou plusieurs profils fréquentiels et caractérisation de la liaison entre les mots considérés.

La première catégorie représente une étude centrée sur la forme des profils fréquentiels qui composent le corpus. Ce type d'étude permet de déduire certaines caractéristiques diachroniques générales de mots en analysant chaque profil fréquentiel indépendamment des autres.

Par exemple, en analysant les profils un à un nous pouvons classer les courbes selon différentes typologies ou bien extraire des caractéristiques simples telles que l'année de fréquence maximale ou la résilience. Par conséquent la visualisation chronocloud fait partie de cette première catégorie.

La deuxième catégorie se réfère à l'étude des relations qu'entretiennent les mots entre eux en fonction de leur profils fréquentiels. L'exemple classique rattaché à cette catégorie est le cas de deux formes "en compétition" désignant le même concept dont l'une viendrait remplacer l'autre, conduisant parfois à l'extinction de la forme initiale. Nous reviendrons sur cette catégorie dans la suite.

Dans la majorité des cas, il est utile de combiner ces deux types de recherches qui, passant systématiquement de l'espace des mots à ceux des profils fréquentiels et vice versa, permettent de repérer des n-grammes et profils fréquentiels particuliers contenant des informations intéressantes et pertinentes pour répondre à diverses questions de recherches préétablies. Ces questions peuvent être d'ordre historique, culturel, sociolinguistique, linguistique ou liées aux études de médias par exemple.

Par conséquent, nous présentons diverses catégories de mots, en premier lieu selon leur sémantique (recherche dans l'espace des mots), mais nous observons également la typologie de courbe de ces mots.

Dimension temporelle

Les années sont la plupart du temps exprimées en une combinaison de quatre chiffres et elle ne sont donc accessibles qu'avec le prétraitement alphanumérique. Le comportement des mots assimilables à des dates est intéressante : ces courbes traduisent un pic attentionnel au moment où l'année correspond à elle même, ce qui est un comportement parfaitement intuitif. Ce qui est peut être moins intuitif, c'est le processus de retour à la normale dans les années qui suivent. Une mémoire subsiste en effet pour chacune de ces années, car le pic est totalement asymétrique. Un exemple des plusieurs profils fréquentiels d'années séparées par 40 ans est présenté dans la Figure 11.68.

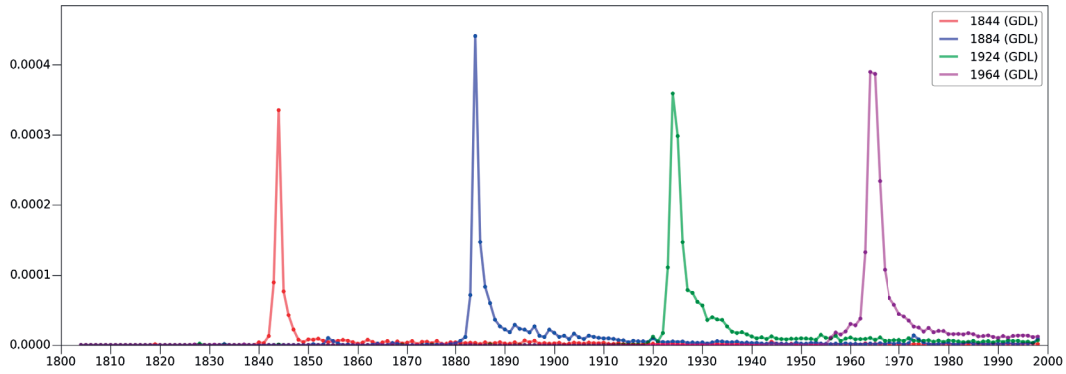


FIGURE 11.68 – Profils fréquentiels de "1844", "1884", "1924" et "1964" dans GDL

Nous observons un comportement similaire entre ces dates, mais le processus de diminution progressive de la fréquence après le pic attentionnel peut être différent d'une date à l'autre. En effet, certaines dates ont acquis une résonance particulière dans notre langue notamment en raison de faits historiques importants comme par exemple les dates de début et fin des deux guerres mondiales. D'autres exemples de dates, mais cette fois ayant un processus de retour à la normale quantitativement différents sont présentés dans la Figure 11.69.

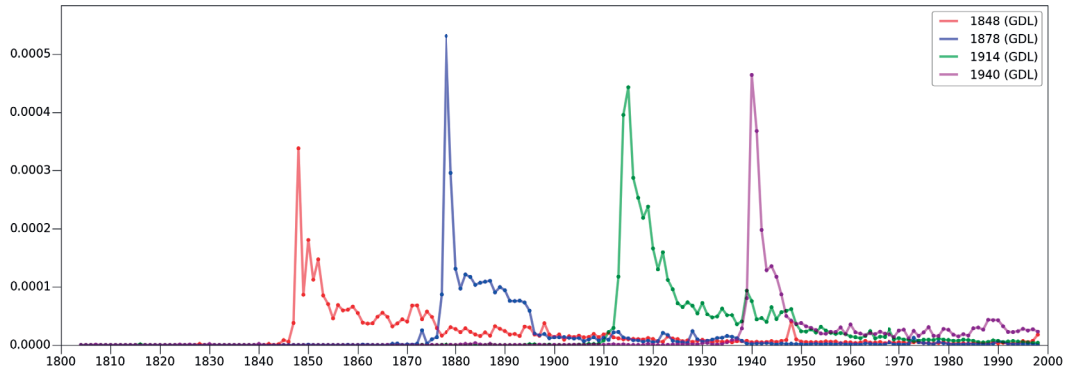


FIGURE 11.69 – Profils fréquentiels de "1848", "1878", "1914" et "1940" dans GDL

Nous observons le régime différent de retour à la normale concernant certaines années particulières, notamment 1914 et 1940 qui sont des années moins rapidement "oubliées" que d'autres. Les différentes périodes historiques (Préhistoire, Antiquité, Moyen-âge ou Renaissance) fournissent d'autres exemples d'indications temporelles. Ces profils fréquentiels sont présentés dans la Figure 11.70.

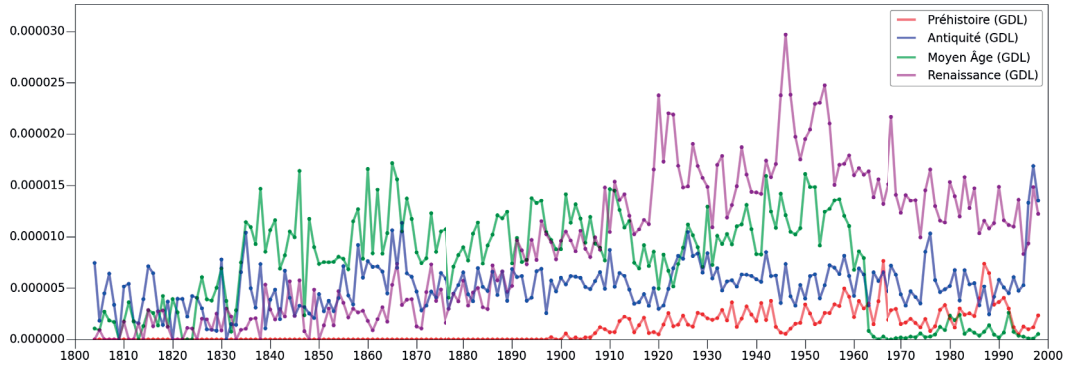


FIGURE 11.70 – Profils fréquentiels de "préhistoire", "antiquité", "moyen-âge" et "renaissance" dans GDL

Nous observons que la fréquence de "renaissance" augmente jusqu'en 1946 avant d'entamer une phase de légère diminution. Le mot "antiquité, potentiellement polysémique, subit une augmentation soudaine en 1996. Le mot "préhistoire" n'apparaît qu'en 1898 et "moyen-age" diminue soudainement en 1963.

Parmi les termes ayant un sens temporel, mais ne désignant pas une date en particulier, on peut citer les mots "passé" et "futur" (cf. Figure 11.71).

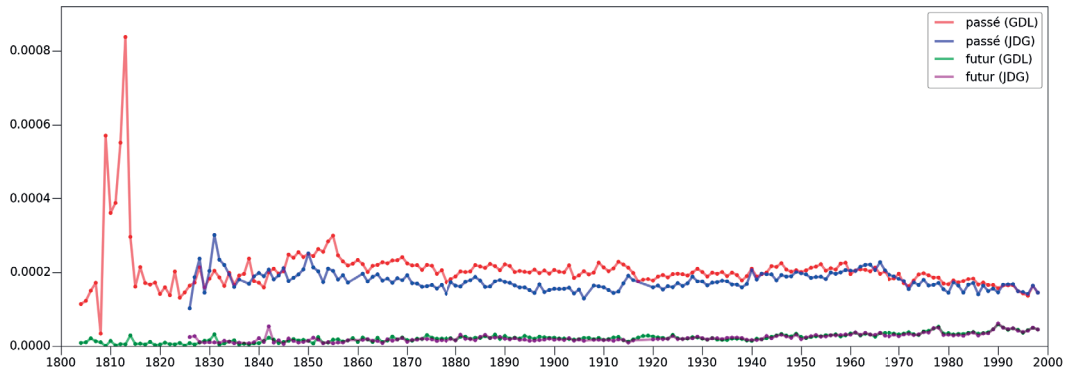


FIGURE 11.71 – Profils fréquentiels de "passé" et "futur"

Ces mots sont beaucoup plus stables que les précédents si ce n'est une légère diminution de "passé" et une légère augmentation de "futur" dans les 30 dernières années du corpus.

Dimension spatiale

Le chronocloud différentiel montre qu'une grande partie des différences de fréquences observées proviennent de mots représentant des localisations géographiques. Ainsi, les pays voisins et leurs capitales et villes ont des profils fréquentiels souvent similaires entre les deux corpus. (cf. Figures 11.72 et 11.73).

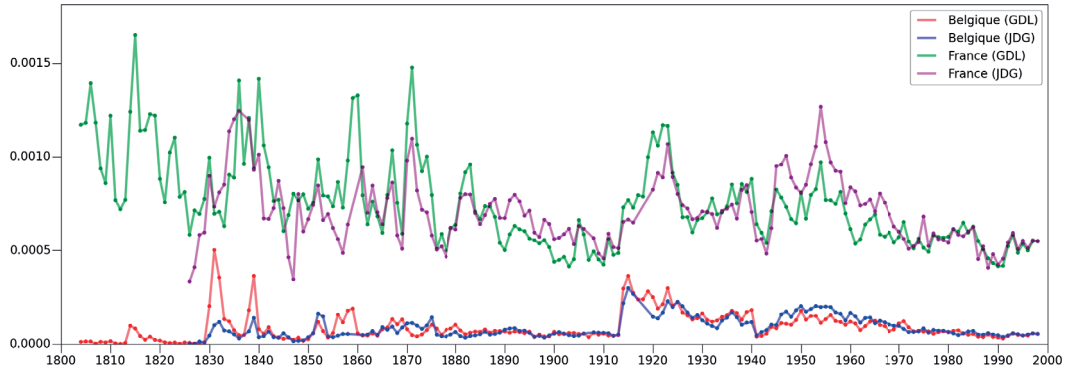


FIGURE 11.72 – Profils fréquentiels de "Belgique" et "France"

Les mentions de "Belgique" et "France" sont similaires dans les deux corpus, mais le profil fréquentiel de "France" est plus élevé que celui de "Belgique". Une hausse de fréquence est constatée pour les deux pays au début de la première guerre mondiale, la hausse étant plus importante pour la Belgique qui était peu mentionnée avant cette guerre. Par après, nous observons un effet de retour à la normale d'après-guerre progressif, la fréquence retournant à peu près à son niveau d'origine.

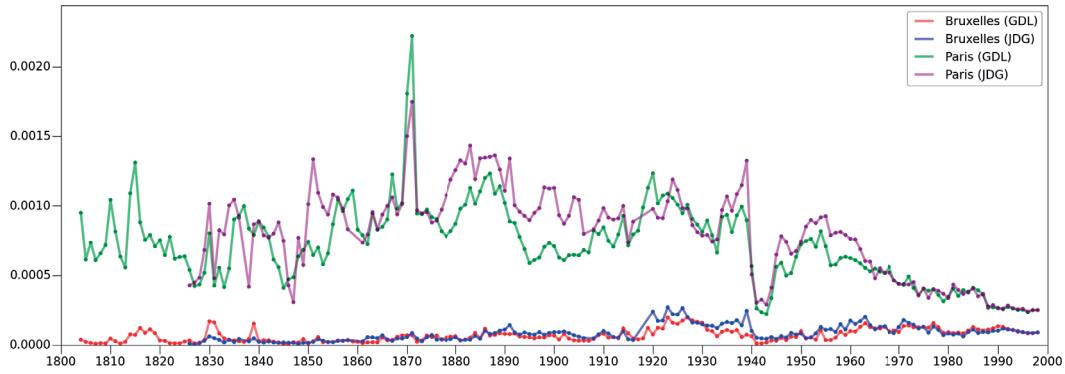


FIGURE 11.73 – Profils fréquentiels de "Bruxelles" et "Paris"

Si l'on se réfère aux capitales "Bruxelles" et "Paris", nous observons un comportement assez similaire aux mots "Belgique" et "France", si ce n'est que "Paris" diminue nettement dans les

deux corpus à partir de 1955. De plus, une chute de fréquence considérable a lieu dès 1940 constatée autant sur les capitales que les pays. Toutefois, nous remarquons que les capitales peuvent être utilisées pour désigner le pays par métonymie. La courbe de "Bruxelles" est plus fortement corrélée avec "Belgique" (0.50 sur JDG et 0.41 sur GDL) et la courbe "Paris" l'est avec "France" (0.41 sur JDG et 0.59 sur GDL). Nous notons également que les courbes "Belgique" et "France" sont corrélées, plus particulièrement dans le corpus de JDG (0.47 sur JDG et 0.32 sur GDL). En allant vers la notion plus globale de continent nous pouvons comparer l'évolution des fréquences de mention des continents au cours du temps (cf. Figure 11.74).

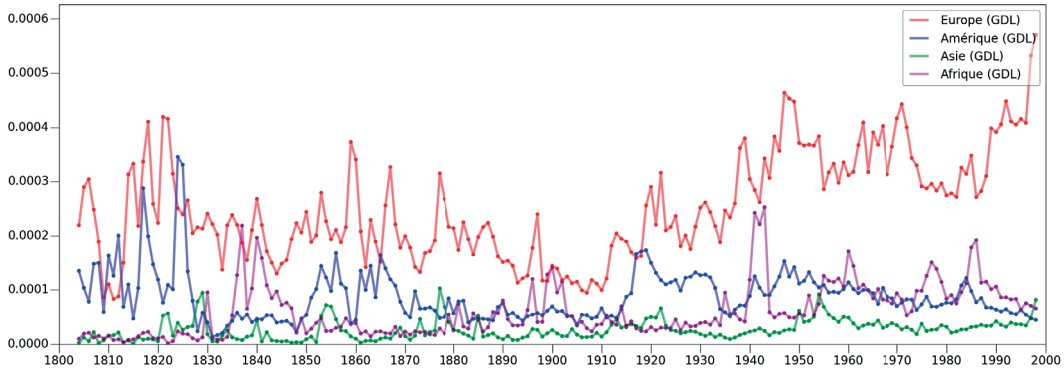


FIGURE 11.74 – Profils fréquentiels de "Europe", "Amérique", "Asie" et "Afrique" dans GDL

Nous observons une augmentation de la fréquence de l'Europe au cours du temps, ce qui est intuitif au vu de sa construction progressive et le fait que la Suisse, bien que non membre de l'Europe, se situe en Europe. D'autres pics de fréquences sont intéressantes au regard de l'étude de l'histoire au travers de ce corpus de presse. En cherchant des mots correspondant à des localisations plus régionales au niveau Suisse, nous trouvons des différences de traitement plus significatives entre les deux journaux. C'est notamment le cas pour les villes de Nyon et Vevey (cf. Figure 11.75).

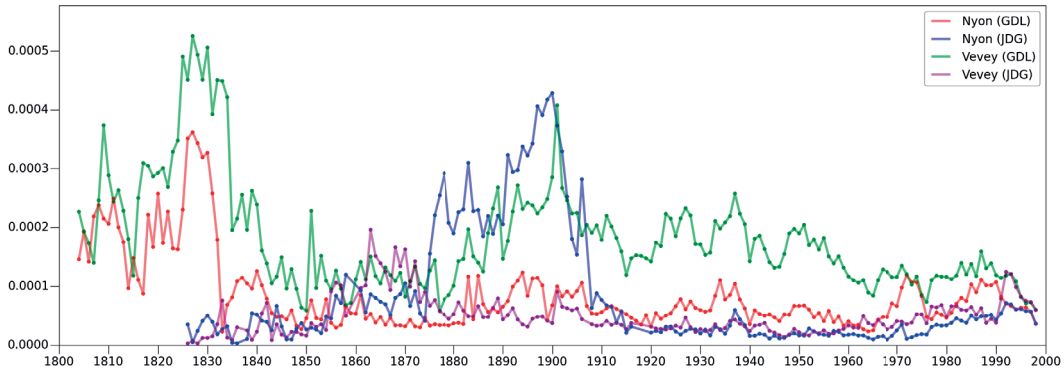


FIGURE 11.75 – Profils fréquentiels de "Nyon" et "Vevey"

Les villes de Nyon et Vevey suivent des courbes différentes, correspondant à la couverture, par les deux journaux des événements liées à ces villes.

Nombres et quantités

Les nombres sont généralement exprimés par des combinaisons de chiffres. Nous pouvons reprendre l'exemple des multiples de 10 afin d'examiner comment la fréquence de la quantité évolue en fonction de la quantité (cf. Figure 11.76).

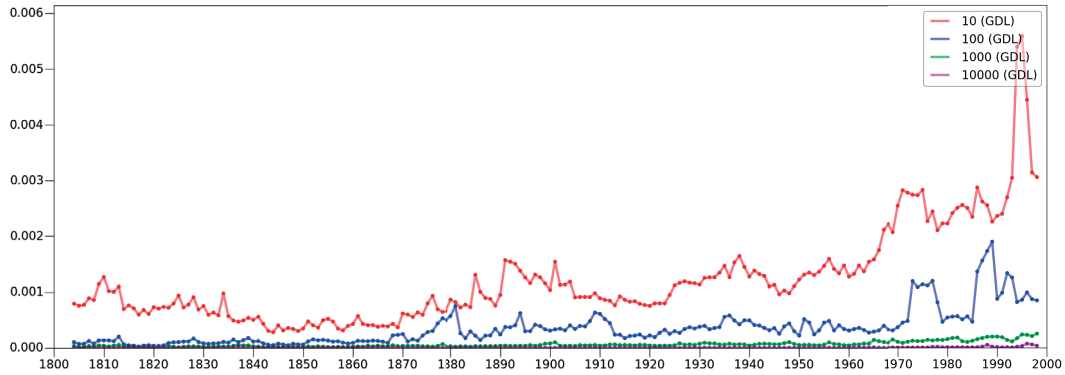


FIGURE 11.76 – Profils fréquentiels de "10", "100", "1000" et "10000" dans GDL

Nous avons déterminé que la principale cause de cette évolution relève de l'introduction de pages de bourse ou page totalement remplie de montants liés à des amortissements financiers. L'analyse diachronique de l'évolution des nombres est donc biaisé par ces sections particulières des journaux GDL et JDG. Toutefois, les nombres s'expriment également en lettres et nous pouvons aussi regarder l'évolution de ces mots comme par exemple les profils fréquentiels de "deux", "quatre" et "huit" (cf. Figure 11.77).

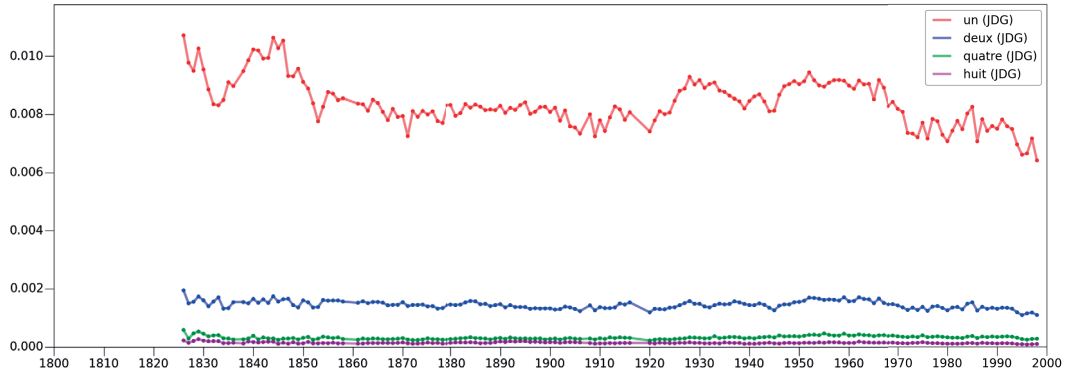


FIGURE 11.77 – Profils fréquentiels de "un", "deux", "quatre" et "huit" dans JDG

Nous remarquons que ces nombres sont plus stables sous la forme de combinaisons de lettres. Le mot "un" apparaît en moyenne 5.73 fois plus que le mot "deux" (5.81 pour JDG et 5.65 pour GDL), le mot "deux" apparaît en moyenne 4.61 fois plus que le mot "quatre" (4.56 pour JDG et 4.66 pour GDL) et le mot "quatre" apparaît en moyenne 2.31 fois plus que le mot "huit" (2.34 pour JDG et 2.29 pour GDL).

D'autres termes de quantité sont utilisés comme les mots "million" et "milliard". Si le mot "million" est plus utilisé que "milliard" en général, nous observons toutefois dans la Figure 11.78 que "milliard" subit une augmentation de fréquence continue et plus accentuée que "million" si bien qu'un croisement a lieu en 1985.

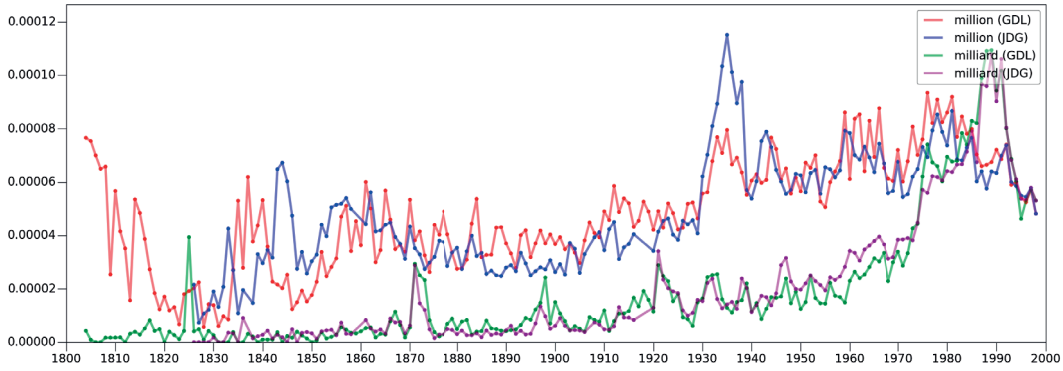


FIGURE 11.78 – Profils fréquentiels de "million" et "milliard"

Si l'on observe les profils fréquentiels des pluriels de ces quantités, comme présenté dans la Figure 11.79, nous constatons un comportement similaire des courbes si ce n'est dans les années les plus récentes où cette fois aucun croisement n'est observé. Il semble clair que le besoin de décrire un niveau de quantité de l'ordre du milliard n'est apparu que tardivement dans ces corpus de journaux, ce qui est d'autant plus vrai pour le pluriel "milliards".

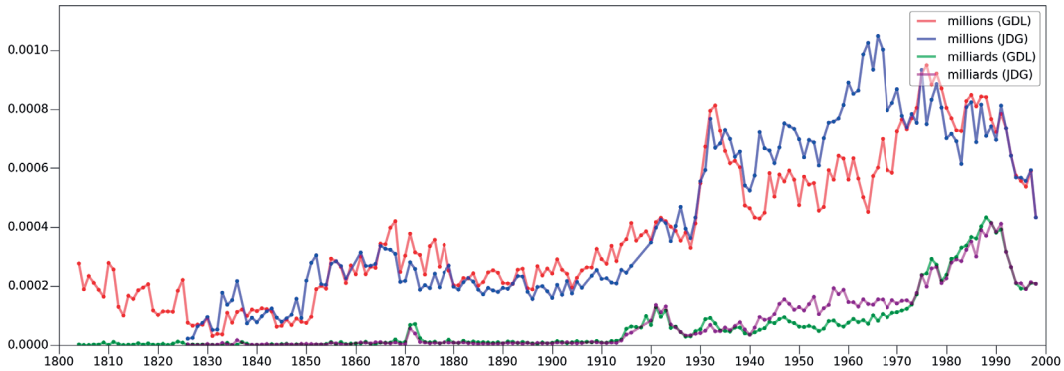


FIGURE 11.79 – Profils fréquentiels de "millions" et "milliards"

En résumé, nous avons observé que les mots représentant des nombres comme des combinaisons de chiffres sont croissants et induits par l'apparition de section boursières dans les journaux. Au contraire, les mots représentant des nombres en toutes lettres sont stables dans le temps et sur les deux corpus. Les mots exprimant de très grande quantités comme "million", "millions", "milliard" et "milliards" sont également en hausse.

Evénements historiques

Nous observons l'impact de grands événements historiques et notamment les deux guerres mondiales. En observant le profil fréquentiel du mot "guerre" (cf. Figure 11.80), nous constatons que la courbe produit deux pics attentionnels significatifs correspondant à la première guerre mondiale (1914-1918) et à la seconde guerre mondiale (1939-1945).

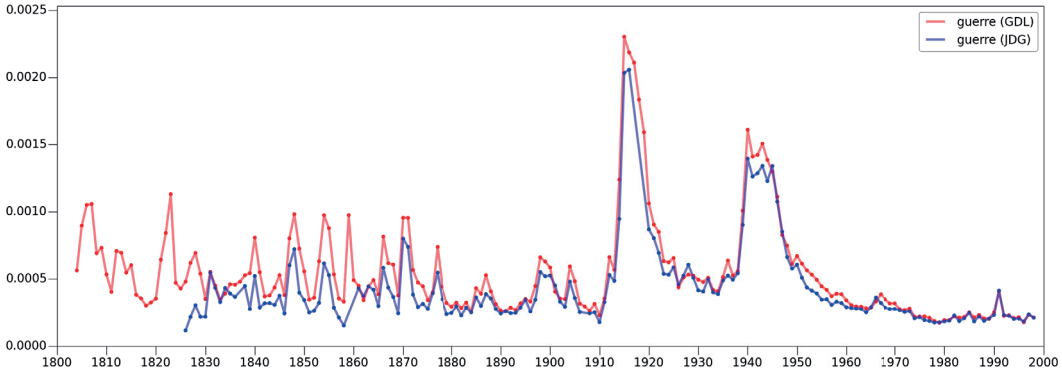


FIGURE 11.80 – Profils fréquentiels de "guerre"

Les mots liés au lexique de la guerre, connaissent également deux pics attentionnels au moment des guerres mondiales comme nous le constatons sur la Figure 11.81 représentant les profils fréquentiels de "guerre", "armée", "soldats" et "bombes".

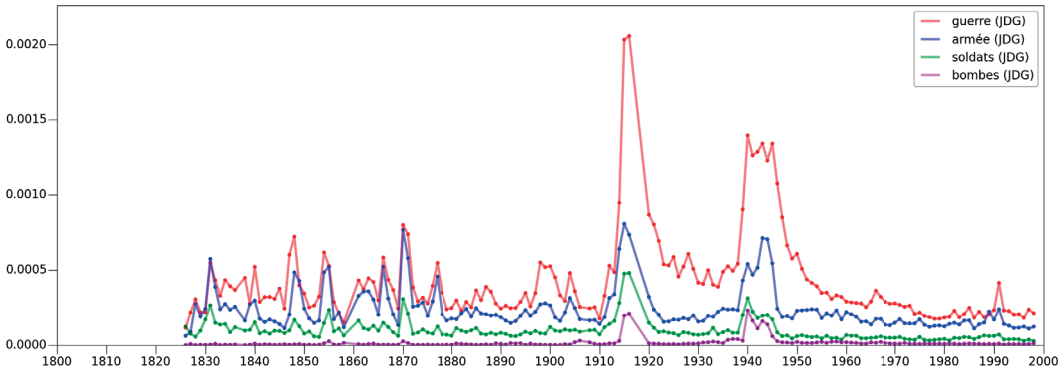


FIGURE 11.81 – Profils fréquentiels de "guerre", "armée", "soldats" et "bombes" dans JDG

En règle générale, les conflits armés sont des événements historiques laissant des traces importantes dans la presse. Les deux guerres mondiales ont impacté le corpus de deux façons distinctes. La première est simplement quantitative, la taille du corpus subit une variation de nombre de mots passant de 9 222 399 en 1913 à 6 871 467 en 1915 soit une diminution de 25.5% et passant de 8 440 260 en 1938 à 5 875 914 en 1940 soit une diminution de 30.4%. La deuxième est qu'elles ont fait entrer à cette époque, dans le vocabulaire usuel, un certain nombre de mots (comme "bombes", "bombardiers", "nazi", "Reich", "Gestapo", "collaboration", "épuration") qui n'étaient pas ou peu utilisés avant les guerres. La question subsiste de savoir si cette impulsion a pu affecter durablement la langue.

L'analyse Macro a montré que l'on peut observer ces effets localement dans le temps (principalement via les notions de distances entre les années et le noyau résilient). Pourtant, la langue de ces deux corpus de journaux ne semble pas être perturbée sur le long terme. Un effet similaire est observé quant aux mots particuliers que nous avons testés (comme "guerre", "armée", "soldats" et "bombes") en lien avec les guerres mondiales. Leurs profils fréquentiels montrent des pics attentionnels qui diminuent ensuite pour un retour à la normale.

Inventions, découvertes et technologies

Ces mots vont généralement s'ajouter au vocabulaire de la langue. Si ces innovations sont de nature à changer le quotidien de l'humanité, alors les mots correspondant deviennent pérennes. Nous abordons dans ce cadre, trois thèmes : l'énergie, les transports et la communication. Dans le domaine de l'énergie nous présentons les profils fréquentiels des mots "électricité", "énergie", "nucléaire" et "solaire" dans la Figure 11.82.

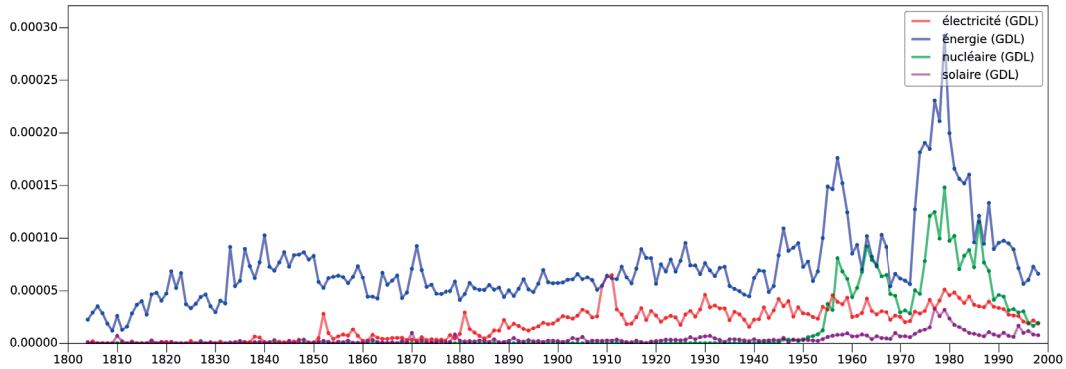


FIGURE 11.82 – Profils fréquentiels de "électricité", "énergie", "nucléaire" et "solaire" dans GDL

Nous remarquons que le mot "nucléaire" est corrélé au mot énergie dès son apparition dans le corpus vers 1950. Corrélée également au mot "nucléaire", le mot "solaire" fait aussi son apparition dans les mêmes années. Il est intéressant de constater que ces trois mots sont liés.

Le mot "électricité" quant à lui apparaît dans le corpus en 1852. Bien qu'en légère diminution dans les années les plus récentes, ces mots restent durablement dans le vocabulaire de la presse. Les nouveaux moyens de transports ont également changé le quotidien des gens et ils sont mentionnés fréquemment dans les journaux. Outre le mot "train" qui est présent tout le long du corpus, nous illustrons les profils fréquentiels des mots "tramway", "tram", "avion" et "bus" dans la Figure 11.83.

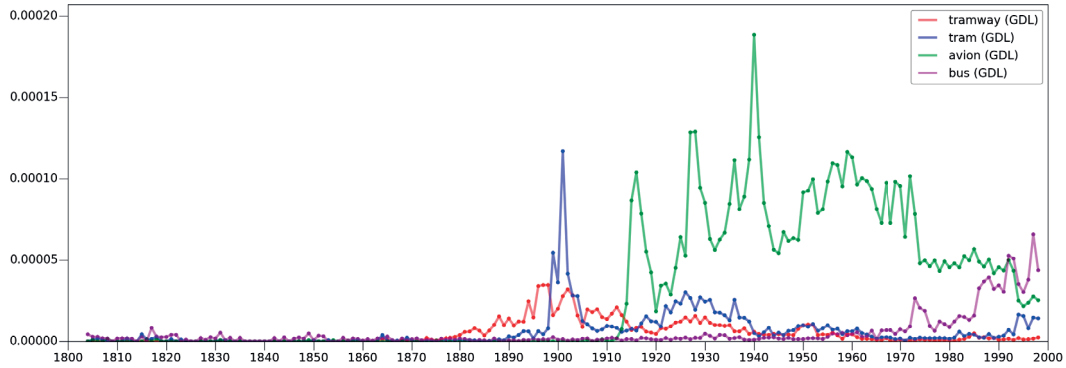


FIGURE 11.83 – Profils fréquentiels de "tramway", "tram", "avion" et "bus" dans GDL

Le mot "tramway" apparaît vers 1880, suivit par "tram" dix ans plus tard, et ensuite "tramway" disparaît tandis que "tram" survit avec une faible fréquence. Le mot "avion" apparaît en 1913 et monte rapidement en fréquence tandis que le mot "bus" apparaît plus tard vers 1956 pour monter en fréquence et finalement croiser la courbe de "avion" en 1992. Hormis "tramway", les autres mots sont restés dans le vocabulaire de la presse.

Enfin, les moyens de communication et les nouvelles technologies prennent des formes de plus en plus diverses au quotidien et dans le langage des êtres humains, y compris dans les corpus de journaux. Les profils fréquentiels des mots "lettre", "télégraphe", "téléphone" et "radio" sont présentés dans les Figures 11.84.

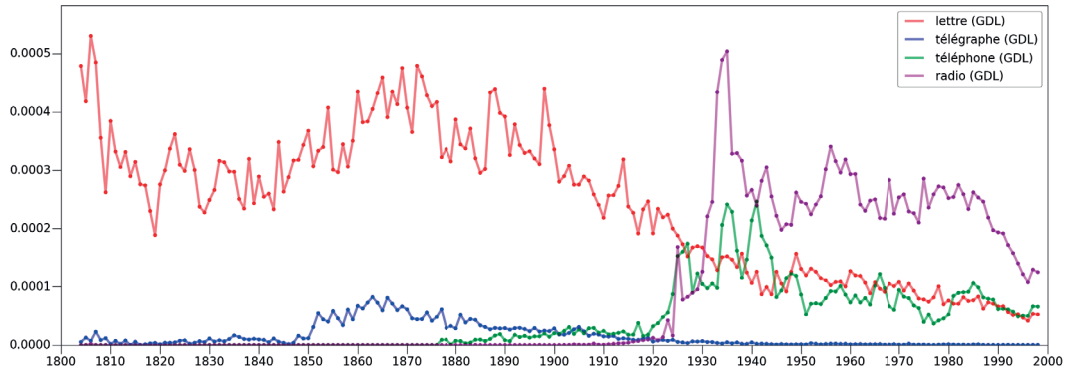


FIGURE 11.84 – Profils fréquentiels de "lettre", "télégraphe", "téléphone" et "radio" dans GDL

L'utilisation du mot "lettre", qui fait partie du noyau résilient, est en constante diminution. Le mot "télégraphe" connaît un pic fréquentiel en 1848, mais garde une fréquence relativement faible par rapport à "lettre" et il s'éteint ensuite vers 1940, dépassé par d'autres technologies.

Le mot "téléphone" est apparu plus tard, en 1877, mais il reste présent dans le corpus jusqu'à la fin. Le mot "radio" apparaît vers 1920, mais sa fréquence augmente rapidement avec le temps, également jusqu'aux dernières années constituant le corpus.

Les profils fréquentiels des mots "télévision", "ordinateur", "informatique" et "internet" sont présentés dans la Figure 11.85.

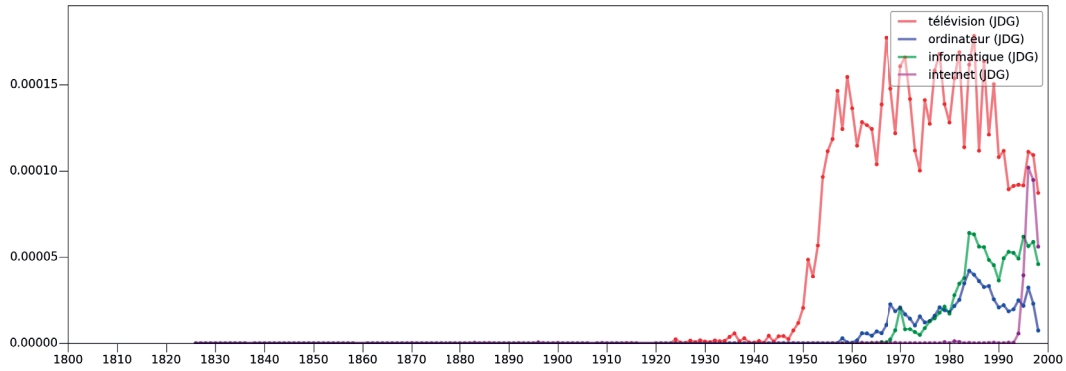


FIGURE 11.85 – Profils fréquentiels de "télévision", "ordinateur", "informatique" et "internet" dans JDG

Bien qu'étant plus récent le mot "télévision" apparaît en 1946 en effectuant une percée importante et se maintient ensuite à une fréquence élevée. Les mots "informatique" et "ordinateur" apparaissent respectivement en 1960 et 1967 avec également un pic fréquentiel important. Le mot "internet" apparaît quant à lui dans les années les plus récentes avec un pic fréquentiel plus élevé que celui de "informatique" et "ordinateur".

En somme, nous avons constaté que les nouveaux moyens de transport, moyens de communication et les nouvelles technologies introduisent de nouveaux mots dans le vocabulaire courant. Ces mots ont également une tendance à effectuer des percées importantes à leur apparition et à maintenir ensuite une fréquence d'utilisation élevée.

Le fait que les nouvelles technologies introduisent de nouveaux mots dans le vocabulaire de la presse écrite n'est pas surprenant, mais il est intéressant de constater que de nombreux mots faisant référence à ces technologies deviennent rapidement fréquents et tendent à se maintenir à ce niveau élevé.

La question de l'apparition de plus en plus fréquente de mots nouveaux dans le vocabulaire, éventuellement lié aux nouvelles technologies, est donc ouverte, mais cela nécessite un corpus qui couvre les années allant au delà de 1998.

Biais de positivité

Les médias et probablement les êtres humains en général, s'expriment avec des mots qui ont des connotations plus positives que négatives. Afin d'illustrer ce biais potentiel, nous présentons les profils fréquentiels des mots "bien" et "mal" dans la Figure 11.86.

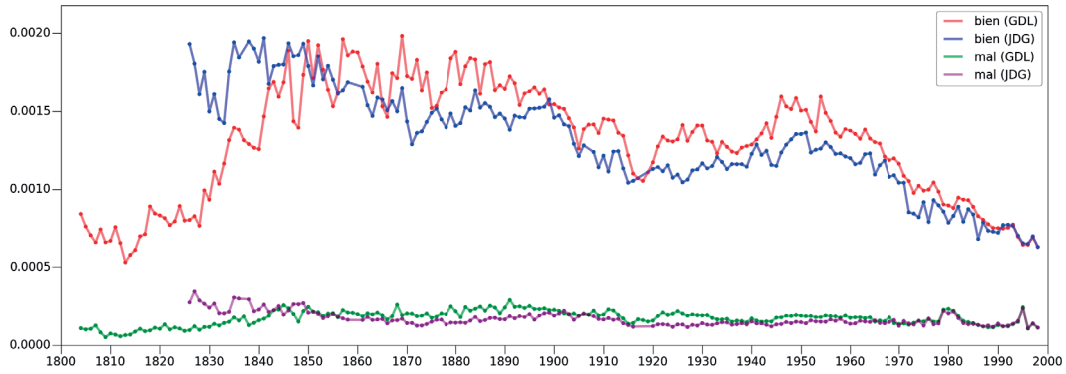


FIGURE 11.86 – Profils fréquentiels de "bien" et "mal"

Nous observons que ces mots suivent une courbe similaire dans les deux corpus. Par contre le mot "bien" possède une fréquence plus élevée que le mot "mal" sur toute la durée du corpus. La corrélation entre les deux mots "bien" et "mal" est de 0.66 dans le corpus de JDG et de 0.72 dans le corpus de GDL.

Toutefois, nous remarquons que le mot bien est polysémique et cela introduit donc un biais supplémentaire. Les profils fréquentiels des mots "bon" et "mauvais" sont présentés dans la Figure 11.87.

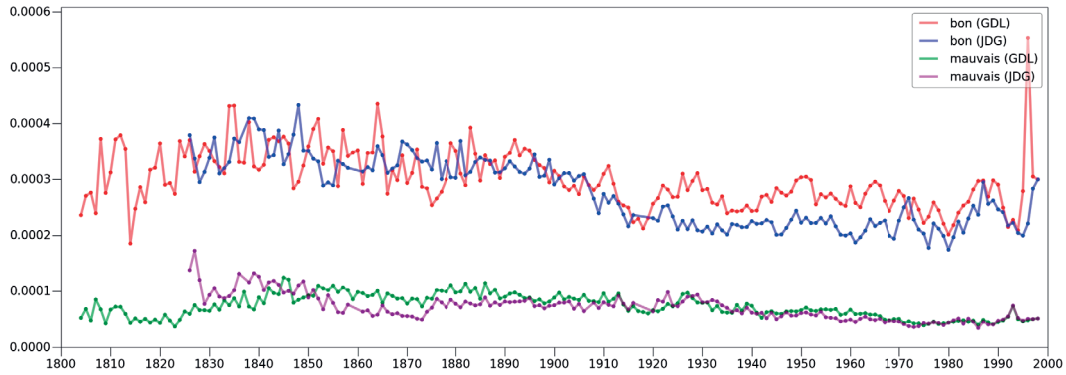


FIGURE 11.87 – Profils fréquentiels de "bon" et "mauvais"

Nous observons que ces mots suivent également une courbe similaire dans les deux corpus. Le mot "bon" possède une fréquence plus élevée que le mot "mauvais" sur toute la durée du

corpus. La corrélation entre les deux mots "bon" et "mauvais" est de 0.61 dans le corpus de JDG et de 0.52 dans le corpus de GDL. Les profils fréquentiels des mots "meilleur" et "pire" sont présentés dans la Figure 11.88.

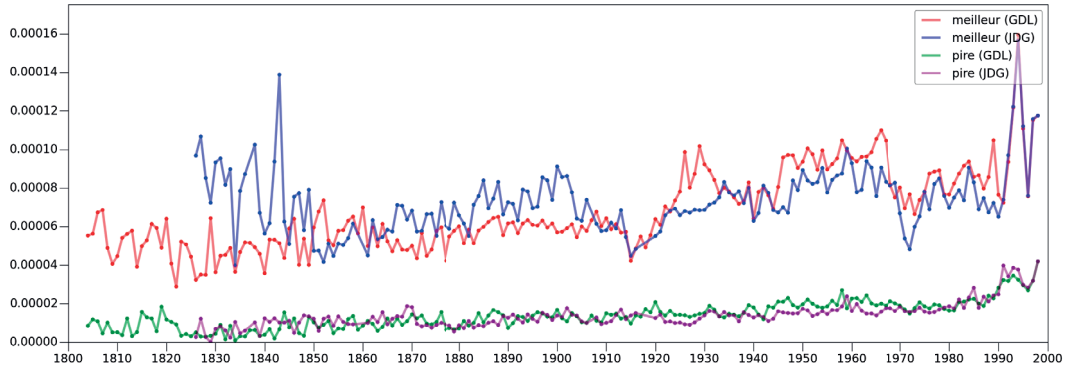


FIGURE 11.88 – Profils fréquentiels de "meilleur" et "pire"

De la même façon, nous observons que ces mots suivent une courbe similaire dans les deux corpus. Encore une fois, le mot "meilleur" possède une fréquence plus élevée que le mot "pire" sur toute la durée du corpus. La corrélation entre les deux mots "meilleur" et "pire" est de 0.43 dans le corpus de JDG et de 0.80 dans le corpus de GDL. les profils fréquentiels des mots "positif" et "négatif" sont présentés dans la Figure 11.89.

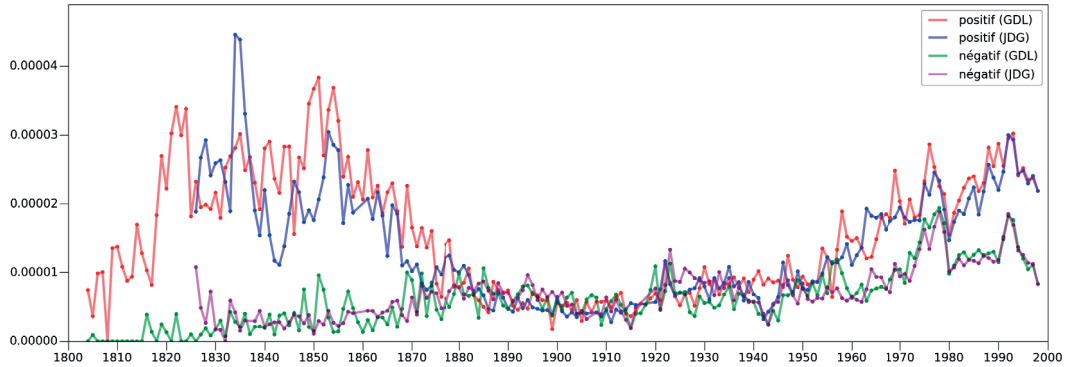


FIGURE 11.89 – Profils fréquentiels de "positif" et "négatif"

Nous remarquons que ces mots suivent aussi une courbe similaire dans les deux corpus. Toutefois, nous observons une tendance variable du mot "positif" qui rejoint le mot "négatif" vers 1880 et s'en détache à nouveau vers 1960. La corrélation entre les deux mots "meilleur" et "pire" est de 0.14 dans JDG et également de 0.14 dans GDL. Malgré ce dernier exemple, nous observons tout de même une tendance générale à ce que la plupart des mots à connotation positive soient plus représentés que leurs correspondants à connotation négative.

Biais de genre

Dans de nombreuse langues, la forme masculine des mots est utilisée lors de généralisation ou de la mise au pluriel. De plus, le contenu du corpus est potentiellement sujet à un tel biais qui est également dépendant du temps, évoluant avec la représentation des genres dans la société. Cela est à mettre en parallèle avec l'évolution du rôle de la femme qui a dû mener de nombreuses batailles pour progressivement conquérir des droits élémentaires. Les profils fréquentiels des mots "homme" et "femme" sont représentés dans la Figure 11.90.

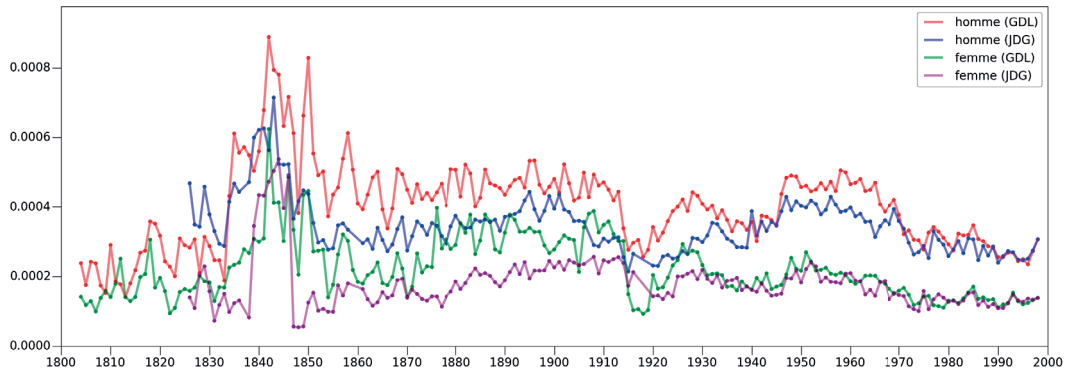


FIGURE 11.90 – Profils fréquentiels de "homme" et "femme"

Les courbes de ces mots affichent des variations élevées et sont fortement perturbées lors de la première guerre mondiale. La corrélation entre les deux mots "homme" et "femme" est de 0.64 dans le corpus de JDG et de 0.80 dans le corpus de GDL. Toutefois, la fréquence moyenne du mot "homme" reste supérieure à celle du mot "femme" même dans les années les plus récentes. Etant donné qu'il est fréquent de désigner tout être humain par "l'homme", y compris la femme, il sans doute plus opportun de regarder une autre forme comme le pluriel. Nous présentons donc les profils fréquentiels des formes plurielles "hommes" et "femmes" dans la Figure 11.91.

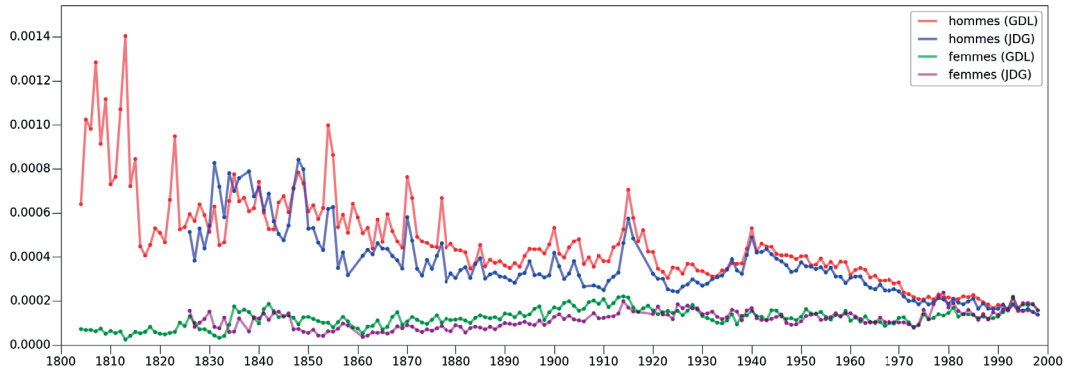


FIGURE 11.91 – Profils fréquentiels de "hommes" et "femmes"

Les courbes des mots "hommes" et "femmes" montrent un comportement différent que celles des mots "homme" et "femme". Leur corrélation est -0.40 dans le corpus de JDG et -0.19 dans le corpus de GDL. Les deux courbes partent dans les années les plus anciennes d'une grande différence de fréquence pour se rejoindre vers 1977. Les profils fréquentiels des mots "un" et "une" avec le prétraitement alpha (car ils sont impactés par le prétraitement alphanumérique) sont représentés dans la Figure 11.92.

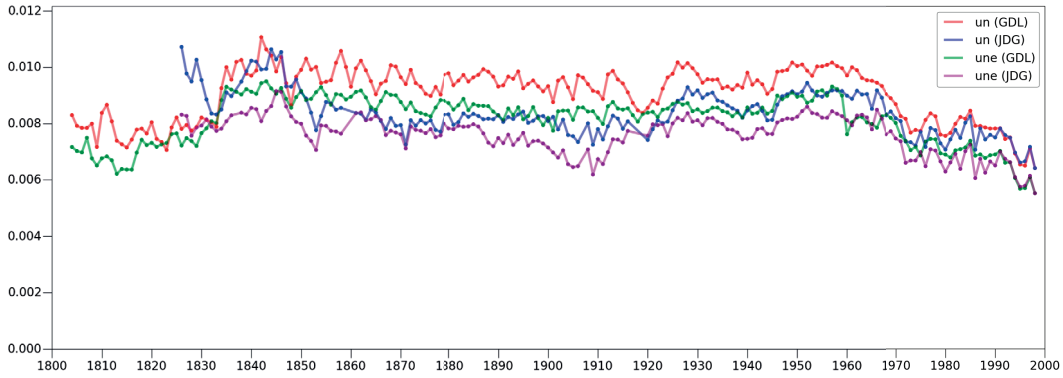


FIGURE 11.92 – Profils fréquentiels de "un" et "une" avec le prétraitement alpha

Nous observons quatre courbes similaires de fréquences élevées. La corrélation entre les deux mots "un" et "une" est de 0.84 dans le corpus de JDG et de 0.93 dans le corpus de GDL. L'ordre fréquentiel des courbes sur la majorité des années est en premier "un" pour GDL, en deuxième et à égalité "un" pour JDG et "une" pour "GDL", enfin en troisième "une" pour JDG.

Dans tous les cas la fréquence du mot "une" reste légèrement inférieure au mot "un". Les profils fréquentiels des mots "il" et "elle" avec le prétraitement alpha (car ils sont également impactés par le prétraitement alphanumérique) sont représentés dans la Figure 11.93.

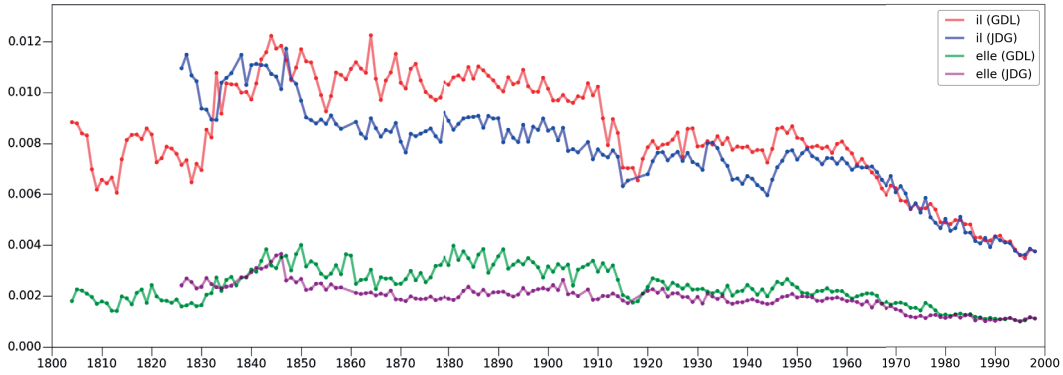


FIGURE 11.93 – Profils fréquentiels de "il" et "elle" avec le prétraitement alpha

Nous observons des courbes similaires dans les deux corpus avec une fréquence nettement supérieur du mot "il" relativement au le mot "elle", mais avec un rapprochement progressif jusque dans les années les plus récentes. Ces quelques exemples de mots en fonction du genre montre que la représentativité des genres dans le langage n'est pas équilibrée, mais qu'un rapprochement s'effectue sur le long terme.

Synthèse

Les analyses de mots particuliers du corpus et leurs profils fréquentiels permettent d'observer des changements linguistiques au niveau des unités de base de la langue. Aucune conclusion généralisée ne peut être affirmée, mais nos observations nous permettent de mieux caractériser le contenu du corpus et les raisons des variations fréquentielles de mots, car nous pouvons à tout moment passer du profil fréquentiel à la recherche directe dans les archives des journaux. Ce type de recherche donne des possibilités d'analyses pouvant emprunter de nombreuses voies de recherche au gré de l'intuition du chercheur et du recoupement de ses diverses recherches précédentes.

Ainsi, chacun de ces profils fréquentiels permet de générer un certain nombre d'hypothèses sur l'évolution du contenu du corpus et de potentiellement généraliser ces hypothèses si le corpus est considéré comme échantillon suffisamment représentatif.

Inversement, il est possible de s'intéresser bien plus aux courbes décrites par les profils fréquentiels qu'aux mots. Faisant l'exercice d'une recherche totalement indépendante du sens du mot, nous constatons que divers types de courbes existent.

Nous présentons les types les plus représentatifs de profils fréquentiels que nous avons observés au travers d'une courte liste :

– Pic attentionnel : ces profils fréquentiels ont un pic attentionnel unique et ensuite s'effacent plus ou moins rapidement (cf. Figure 11.94).

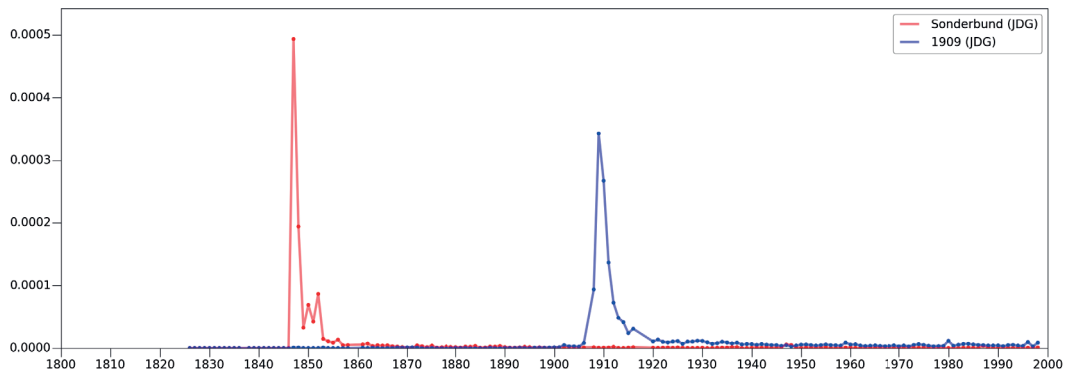


FIGURE 11.94 – Profils fréquentiels de "Sonderbund" et "1909" dans JDG

– Peigne : ces profils fréquentiels ont un pic attentionel cyclique qui se répète de façon périodique dès l'apparition des mots dans le corpus (cf. Figure 11.95).

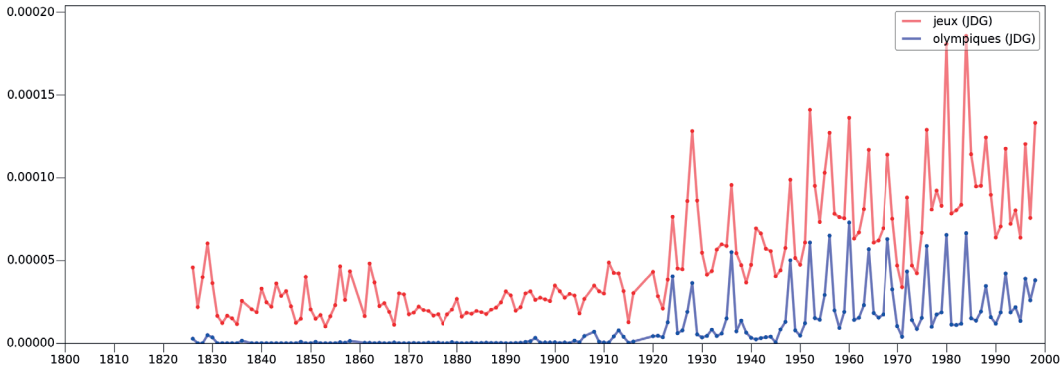


FIGURE 11.95 – Profils fréquentiels de "jeux" et "olympiques" dans JDG

– Décroissance : ces profils fréquentiels voient leur fréquence diminuer le long du corpus.

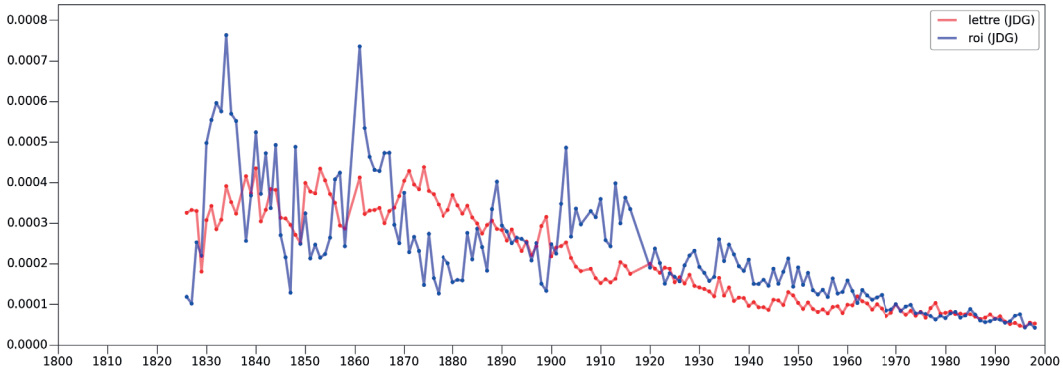


FIGURE 11.96 – Profils fréquentiels de "lettre" et "roi" dans JDG

– Croissance : ces profils fréquentiels voient leur fréquence augmenter le long du corpus.

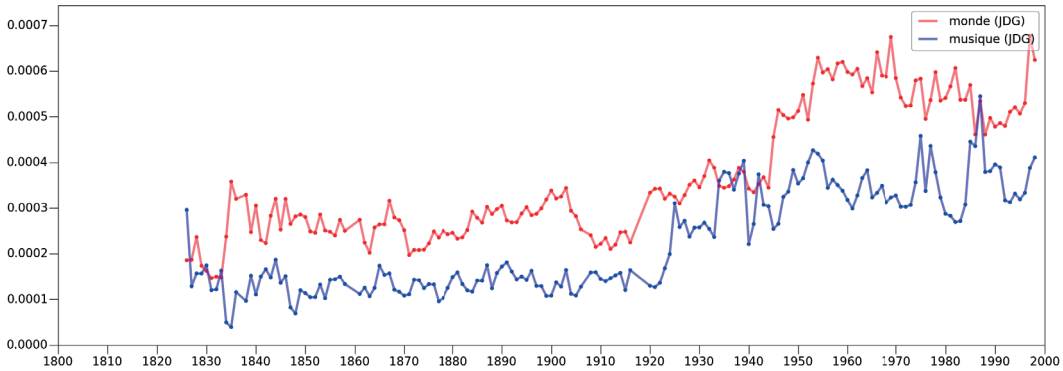


FIGURE 11.97 – Profils fréquentiels de "monde" et "musique" dans JDG

– Disparition d'anciens mots :

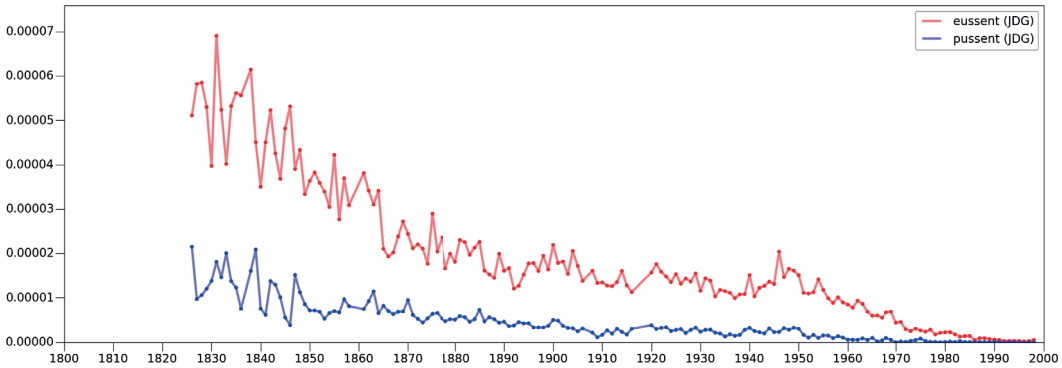


FIGURE 11.98 – Profils fréquentiels de "eussent" et "pussent" dans JDG

– Apparition de nouveaux mots :

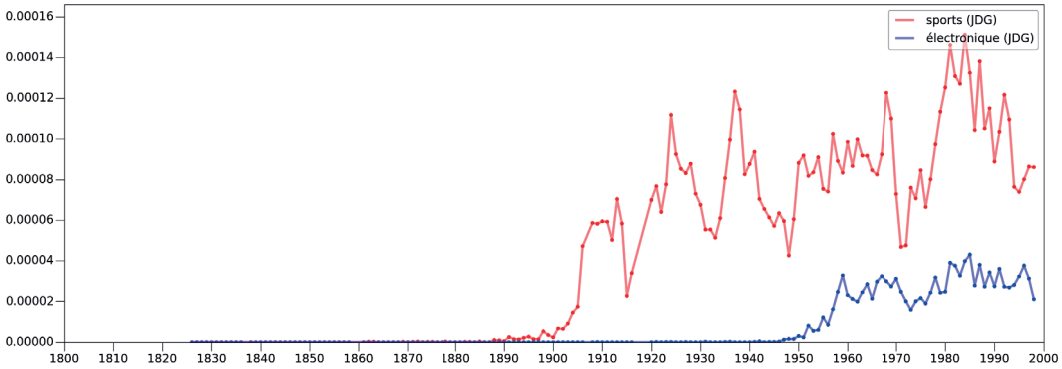


FIGURE 11.99 – Profils fréquentiels de "sports" et "électronique" dans JDG

– Apparition et disparition de mots :

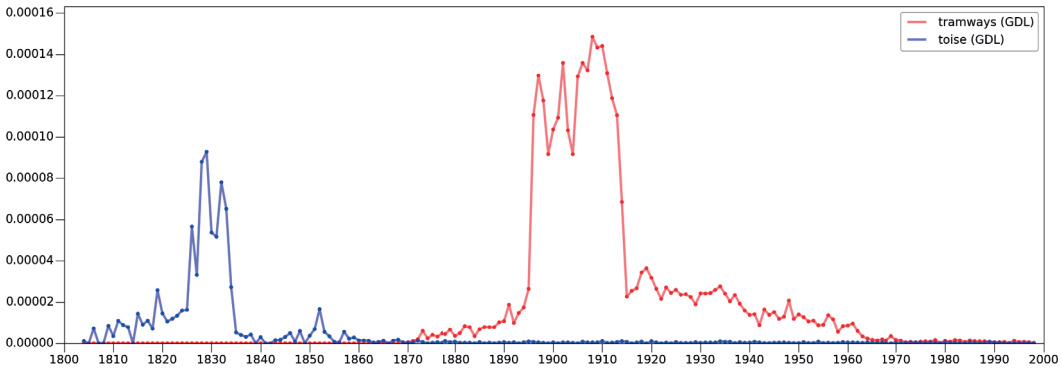


FIGURE 11.100 – Profils fréquentiels de "tramways" et "toise" dans GDL

– Une colline

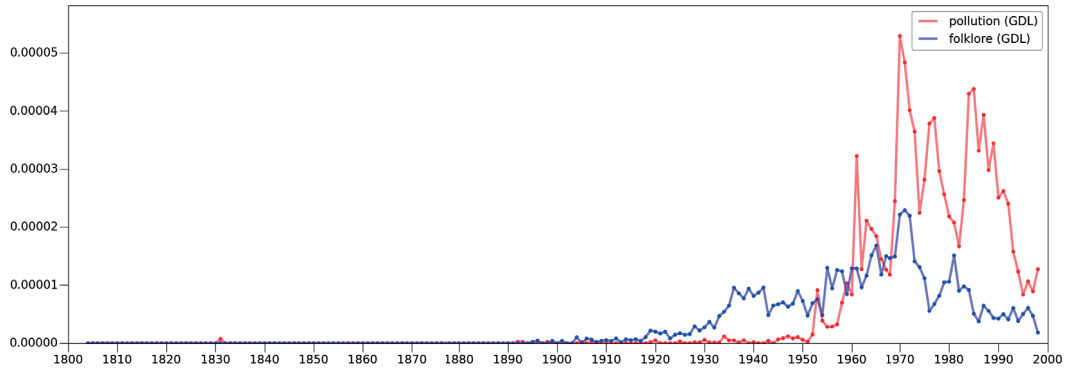


FIGURE 11.101 – Profils fréquentiels de "pollution" et "folklore" dans GDL

D'autres phénomènes méritent une attention particulière comme ceux qui se reflètent dans les relations entre deux courbes. Ces phénomènes ne sont donc détectables qu'en étudiant les courbes deux à deux au minimum.

Un exemple de courbes qui se rejoignent est présenté dans la Figure 11.102.

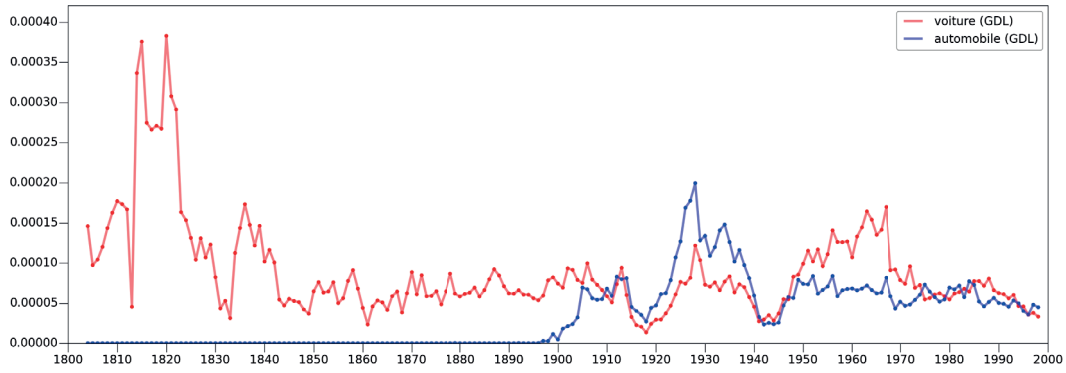


FIGURE 11.102 – Profils fréquentiels de "voiture" et "automobile" dans GDL

Cet exemple montre que le mot "automobile" apparaît au début de XXe siècle et est immédiatement proche de mot "voiture". Le sens de "voiture" est plus général mais il est devenu populaire d'appeler une automobile par le terme "voiture".

Un autre phénomène intéressant est le remplacement d'une courbe par une autre. Cette dernière catégorie montre le processus qui attache le sens d'une forme donnée à une autre pour des raisons d'usages ou simplement des réformes orthographiques.

Un exemple de courbe qui en remplace une autre est présenté dans la Figure 11.103.

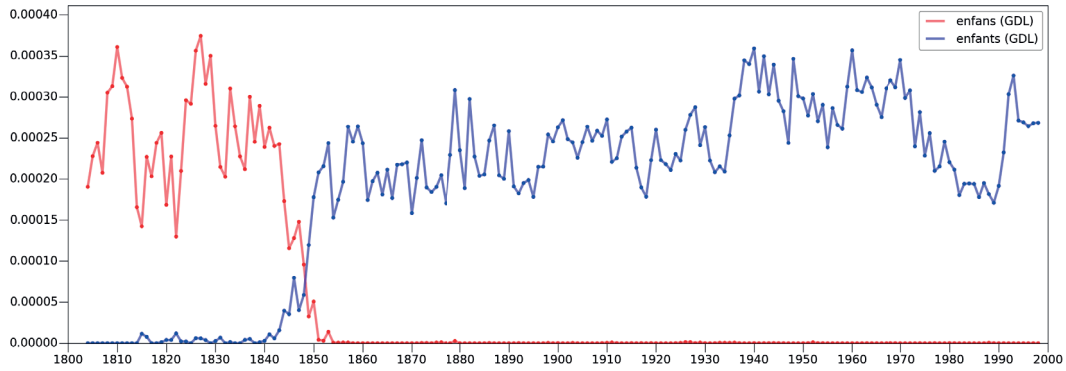


FIGURE 11.103 – Profils fréquentiels de "enfans" et "enfants" dans GDL

Ce remplacement est le résultat d'une réforme orthographique française de l'année 1835, changeant entre autre de nombreux mots qui au pluriel se terminent par "ns" pour qu'ils se terminent par "nts".

Un autre exemple du même type est montré dans la Figure 11.104 avec les mots "violens" et "violents".

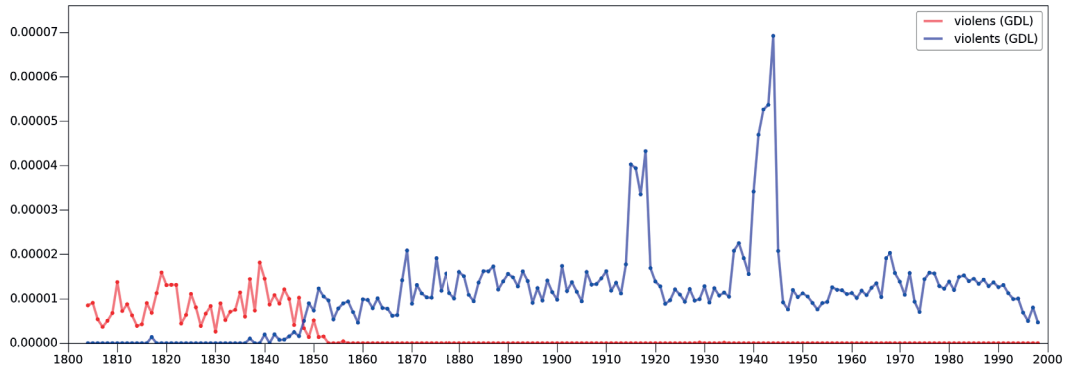


FIGURE 11.104 – Profils fréquentiels de "violens" et "violents" dans GDL

C'est donc le même comportement des courbes qui sera observé sur la plupart des exemples touchés par la reforme orthographique de 1835 "agens" et "agents", "momens" et "moments", "parens" et "parents", "récens" et "récents", "élémens" et "éléments", "accidens" et "accidents", "adhérens" et "adhérents", "innocens" et "innocents", "résidens" et "résidents", "incidens" et "incidents", "placemens" et "placements", etc.

Un dernier exemple de conséquence de la réforme orthographique est montré dans la Figure 11.105 avec les mots "budget" et "budjet".

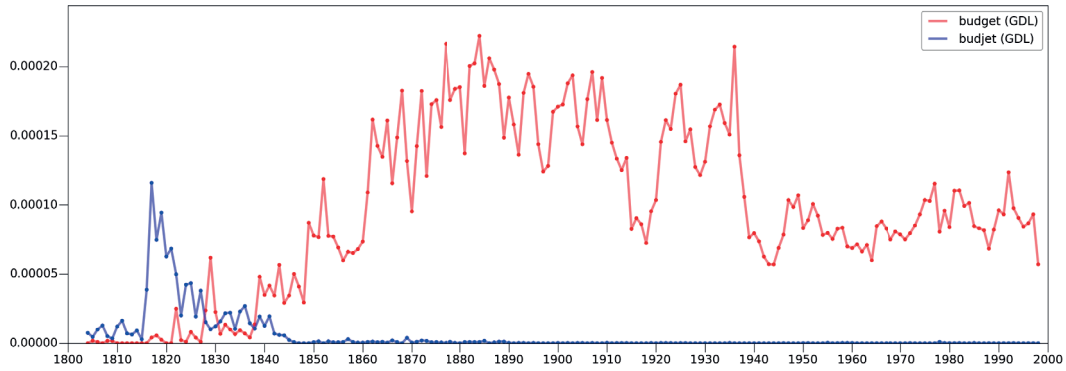


FIGURE 11.105 – Profils fréquentiels de "budget" et "budjet" pour GDL

Nous avions également présenté ce type de processus dans l'analyse des chronoclouds avec les mots "wagon" et "vagon", "Shanghaï" et "Changhaï", "Tokio" et "Tokyo" ainsi qu'avec "tsar" et "czar" (cf. Figures 11.58, 11.60 et 11.64) montrant que les deux journaux ne sont pas toujours en phase sur ce type de changement, car ils interviennent parfois à des moments différents ou même dans certains cas, le changement n'aura pas lieu dans l'autre journal. Enfin, nous avons constaté l'apparition de nombreux mots dans les années les plus récentes.

Certains de ces mots sont liés à des évolutions de société et d'autres sont issus des nouvelles technologies. Les profils fréquentiels des mots "racisme", "écologie", "homosexuel" et "web" sont présentés dans la Figure 11.106.

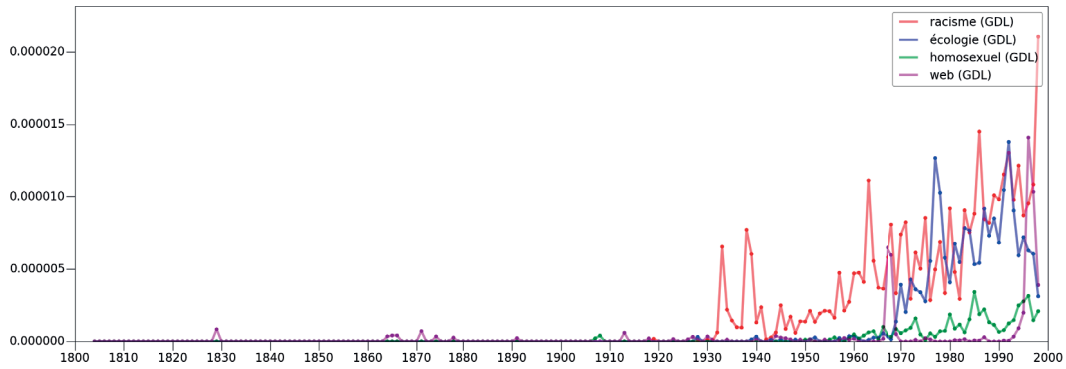


FIGURE 11.106 – Profils fréquentiels de "racisme", "écologie", "homosexuel" et "web" pour GDL

La plupart de ces nouveaux mots ont encore une fréquence faible à la fin du corpus, mais ils entament une phase d'augmentation dont la stabilisation n'est pas connue et peuvent jouer un rôle clé dans les analyses de corpus couvrant ces dates plus récentes. Il est donc intéressant de constater leur présence dans la fin de ces corpus.

12 Analyse de (2-9)-grammes

Dans ce chapitre, nous analysons les n-grammes de niveaux $n > 1$. Nous quittons donc l'espace des mots pour entrer dans l'espace des combinaisons de n mots consécutifs. L'objectif est d'utiliser les concepts et outils développés dans les chapitres précédents pour analyser l'évolution linguistique diachronique des n-grammes en caractérisant potentiellement les comportements des différents niveaux n d'analyse. Nous commençons, comme dans le chapitre précédent, par l'analyse diachronique des distances. Ensuite nous analyserons l'évolution de l'entropie du système ainsi que l'entropie nucléaire, entropie du noyau résilient. Nous terminons par l'étude du niveau Micro via les chronoclouds classiques et différentiels en synergie avec le visualisateur de n-grammes.

12.1 Analyse diachronique des distances

En préambule à l'analyse diachronique des distances, il est nécessaire de préciser les données relatives à la taille des ensembles considérés. Pour le calcul de la distance de Jaccard, nous avions sélectionné les mots dont la fréquence dépassait un seuil de 1 sur 100 000. Etant donné le grand nombre de mots de vocabulaire dont la fréquence se situe en dessous de cette limite (queue de distribution de type zipfienne), la proportion d'éléments uniques filtrés avoisine les 90%. En même temps, la distance de Jaccard étant sensible aux différences de taille du corpus, le filtre fréquentiel appliqué permet essentiellement de réduire cette sensibilité bien qu'elle soit toujours présente, tout en réduisant également l'impact de diverses erreurs d'OCR sur la liste de mots uniques. En effet, si dans notre cas les erreurs d'OCR ont un impact potentiellement faible sur le total des mots, il n'en est pas de même (de par leur caractère aléatoire) sur la liste des mots uniques. En augmentant le niveau n des n-grammes, ces erreurs auront encore plus d'impact sur la liste des n-grammes uniques utilisée dans le cadre du calcul de la distance de Jaccard et de la détermination du noyau résilient.

De plus, un effet supplémentaire à considérer est le fait que plus un n-gramme est long et plus sa fréquence est faible en moyenne. En effet, à des valeur de n comme 9 et plus, le n-gramme devient quasiment une phrase et sa répétition dans le corpus est alors plutôt le signe d'un

copier-coller que d'une véritable création individuelle de suite de mots. Cet effet est mesurable, en comptant le nombre de hapax par année, i.e. le nombre de n-grammes qui n'apparaissent qu'une seule fois sur l'année, en rapport avec le nombre de n-gramme total de cette année et ce en fonction de n. L'évolution de la moyenne de la proportion de hapax en fonction de n est présentée dans la Figure 12.1.

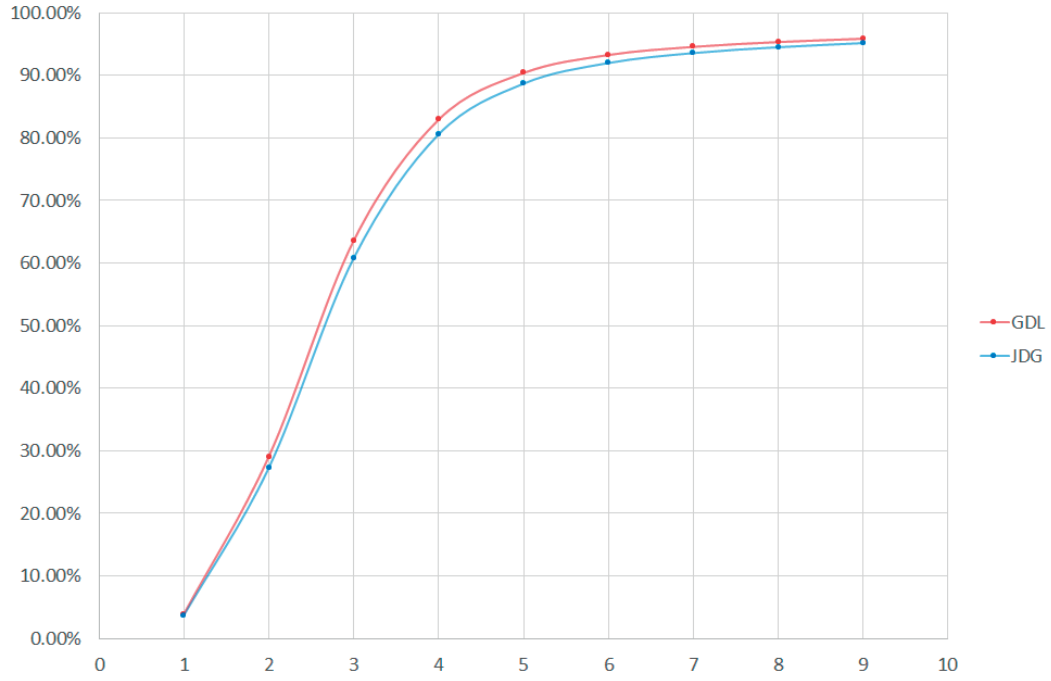


FIGURE 12.1 – Proportion moyenne de hapax par année en fonction de n

Nous observons à $n = 1$ une proportion moyenne de hapax de 3.81% pour GDL et de 3.76% pour JDG. Il est clair que ces mots n'ont que peu d'impacts en terme de fréquence en raison d'une apparition minimale et non reproduite sur le corpus. Ces mêmes proportions atteignent rapidement un taux supérieur à 90% dès le niveau $n = 6$.

Par conséquent, le nombre de n-grammes ayant plus d'une occurrence a tendance à être également plus faible avec le niveau n. Le filtre fréquentiel aura donc un effet plus puissant. Nous avons effectué les mêmes calculs avec les filtres suivants : fréquence supérieure à 1 sur 100 000, fréquence supérieure à 1 sur 1 000 000 et nombre d'occurences supérieur à 1 (filtre minimal). Il en ressort différents niveaux de lissage de la matrice de distance de Jaccard, mais les tendances sont similaires.

L'analyse des distances de Jaccard munie du filtre fréquentiel de base (seuil fréquentiel fixé à 1 sur 100 000) ne peut donc pas être appliqué à des niveaux $n > 5$, car la distance ne distingue plus ou peu les différents sous-corpus annuels et, quant elle le fait, elle montre une variabilité trop importante.

De façon identique, l'exploitation de la notion de noyau résilient ne peut se faire que sur un niveau $n < 6$ également. En effet, dans cette méthode la limitation se situe autour de la détermination du noyau résilient qui a tendance à se réduire avec le niveau n (sauf pour $n = 2$). Dans le cas de notre corpus, le noyau résilient ne contient plus aucun n-grammes dès le niveau $n = 6$. Le nombre de n-grammes composant le noyaux résilient en fonction de n est présenté dans la Figure 12.2.

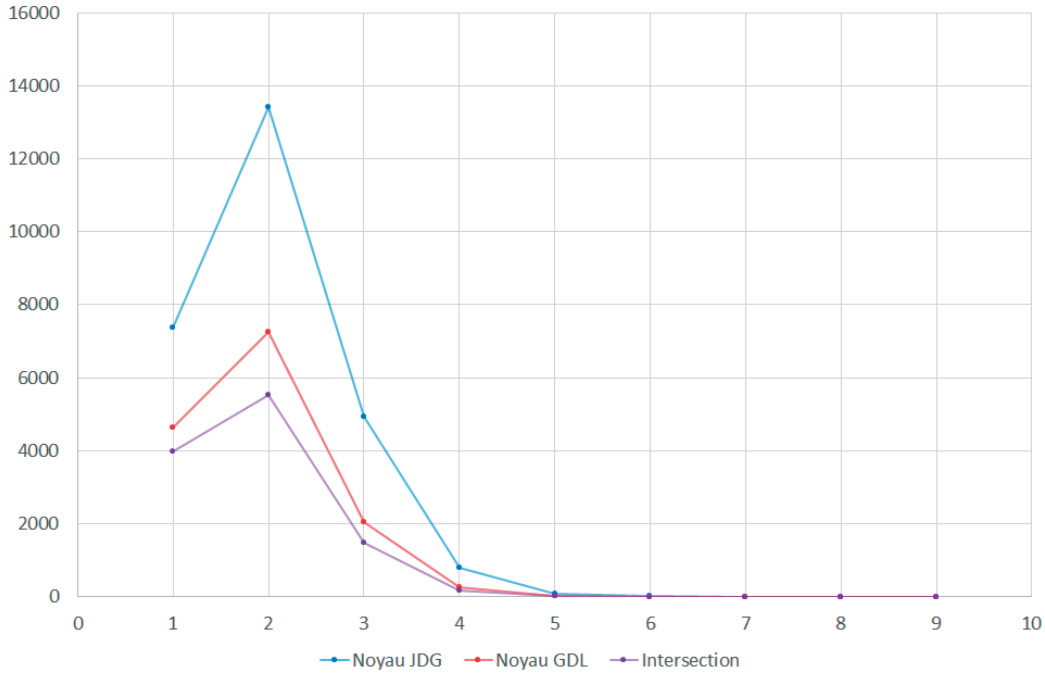


FIGURE 12.2 – Nombre de n-grammes composant le noyau résilient en fonction de n

Nous observons que le noyau résilient augmente au niveau $n = 2$ et diminue ensuite jusqu'à atteindre la valeur nulle. Nous remarquons que la tendance est la même pour les deux corpus. Le nombre absolu d'éléments composant le noyau résilient dépend de la période couverte par le corpus étudié qui est différente dans les deux journaux. La taille du noyau résilient commun aux deux corpus est donc essentiellement limité par le corpus couvrant une plus longue période, dans notre cas le corpus de GDL.

Il est intéressant de constater que le noyau résilient, notion mise en place afin de sélectionner les n-grammes qui ont un rôle potentiellement plus important dans le fonctionnement de la langue de par leur stabilité temporelle, disparaît dans les deux cas vers $n = 5$.

Au délà de ces considérations limitant les niveaux étudiés à $n < 6$, l'analyse des distances de Jaccard et nucléaires entre les sous-corpus annuels est reproductible sans modification conceptuelle au niveau de n-grammes. Les résultats de ces méthodes sur les niveaux supérieurs sont présentés par niveau dans les Figures 12.3, 12.4, 12.5 et 12.6.

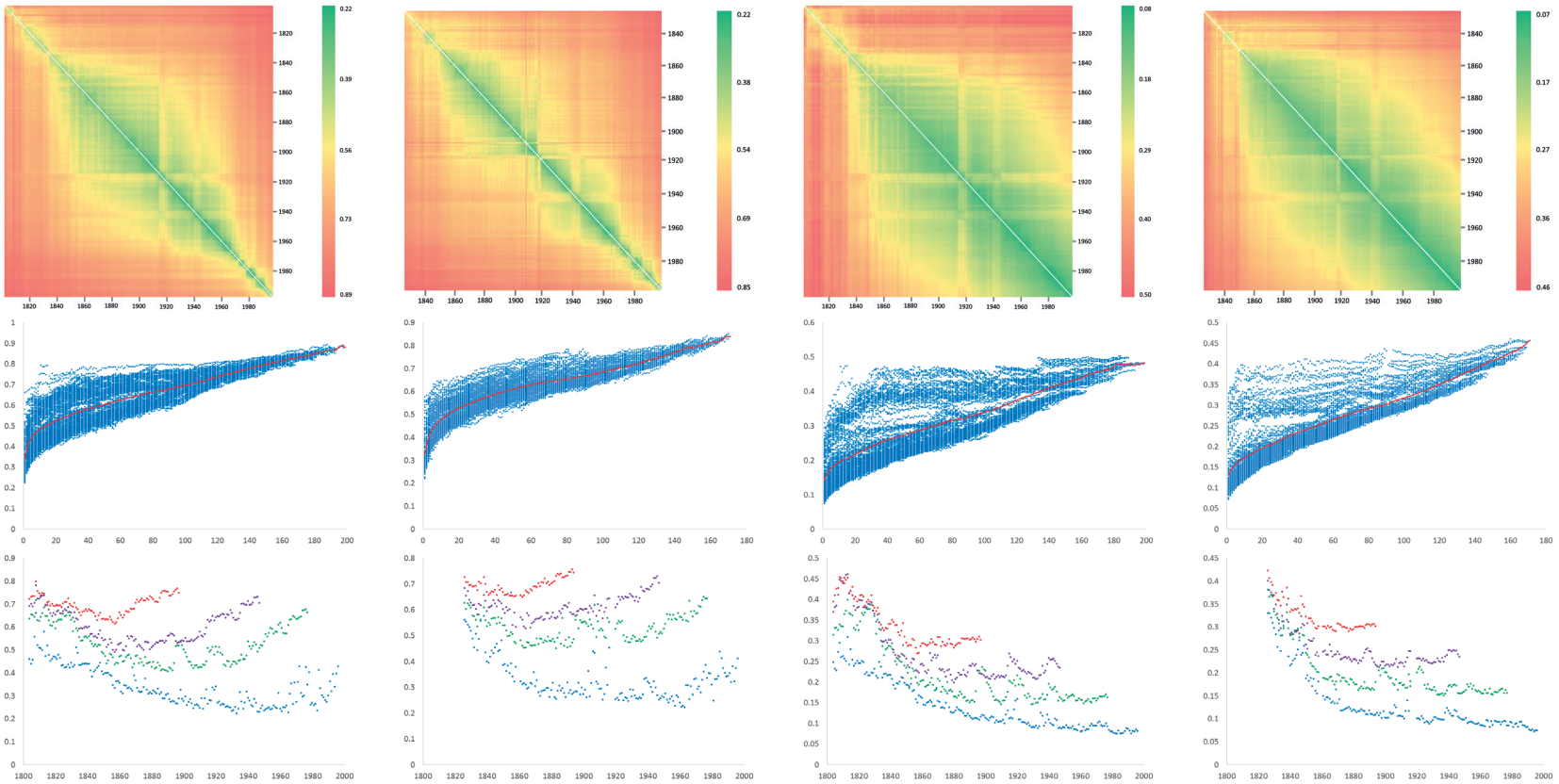


FIGURE 12.3 – **(1)** : Distances de Jaccard sur les 2-grammes de GDL ; **(2)** : Distances de Jaccard sur les 2-grammes de JDG ; **(3)** : Distances nucléaires sur les 2-grammes de GDL ; **(4)** : Distances nucléaires sur les 2-grammes de JDG ; **Haut** : Heatmap de la matrice des distances ; **Milieu** : Distances (bleu) et moyenne de ces distances (rouge) en fonction du nombre d'années de différence entre les sous-corpus ; **Bas** : Distances entre les années y_i et y_{i+n} avec $n = 1$ (bleu), $n = 20$ (vert), $n = 50$ (violet) and $n = 100$ (rouge)

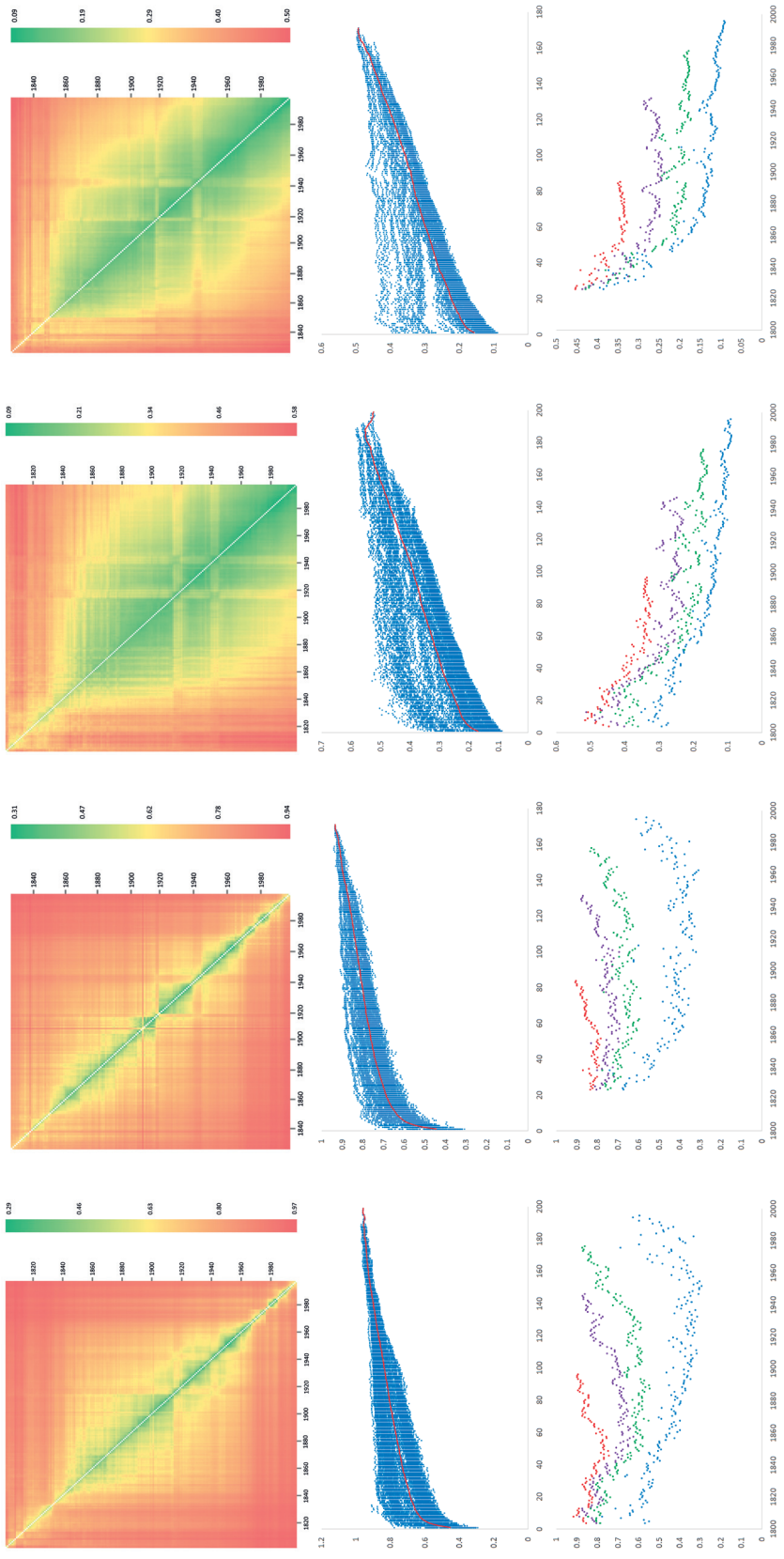


FIGURE 12.4 – **(1)** : Distances de Jaccard sur les 3-grammes de GDL ; **(2)** : Distances de Jaccard sur les 3-grammes de JDG ; **(3)** : Distances nucléaires sur les 3-grammes de GDL ; **(4)** : Distances nucléaires sur les 3-grammes de JDG ; **Haut** : Heatmap de la matrice des distances ; **Milieu** : Distances (bleu) et moyenne de ces distances (rouge) en fonction du nombre d'années de différence entre les sous-corpus ; **Bas** : Distances entre les années y_i et y_{i+n} avec $n = 1$ (bleu), $n = 20$ (vert), $n = 50$ (violet) and $n = 100$ (rouge)

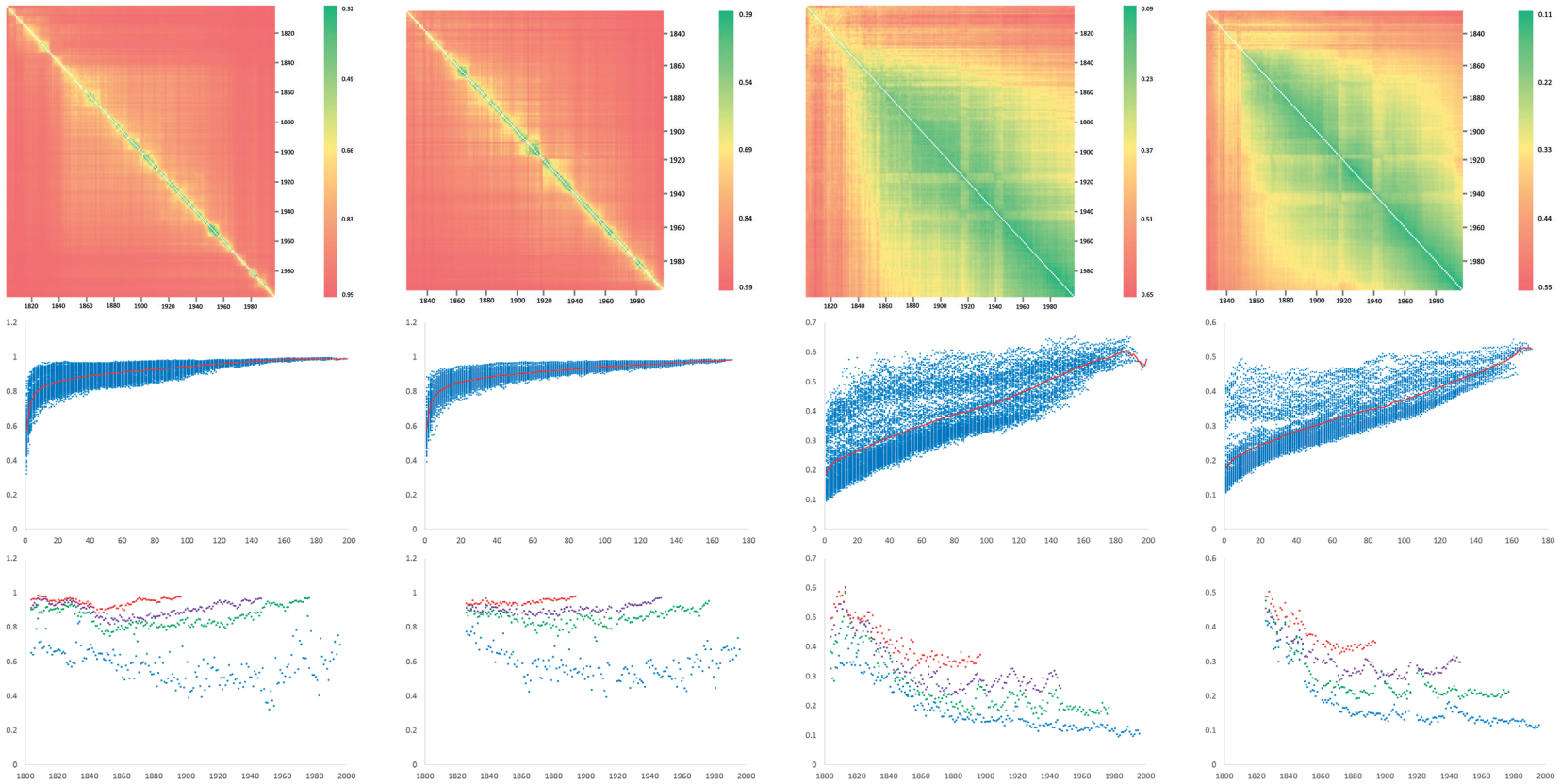


FIGURE 12.5 – **(1)** : Distances de Jaccard sur les 4-grammes de GDL ; **(2)** : Distances de Jaccard sur les 4-grammes de JDG ; **(3)** : Distances nucléaires sur les 4-grammes de GDL ; **(4)** : Distances nucléaires sur les 4-grammes de JDG ; **Haut** : Heatmap de la matrice des distances ; **Milieu** : Distances (bleu) et moyenne de ces distances (rouge) en fonction du nombre d'années de différence entre les sous-corpus ; **Bas** : Distances entre les années y_i et y_{i+n} avec $n = 1$ (bleu), $n = 20$ (vert), $n = 50$ (violet) and $n = 100$ (rouge)

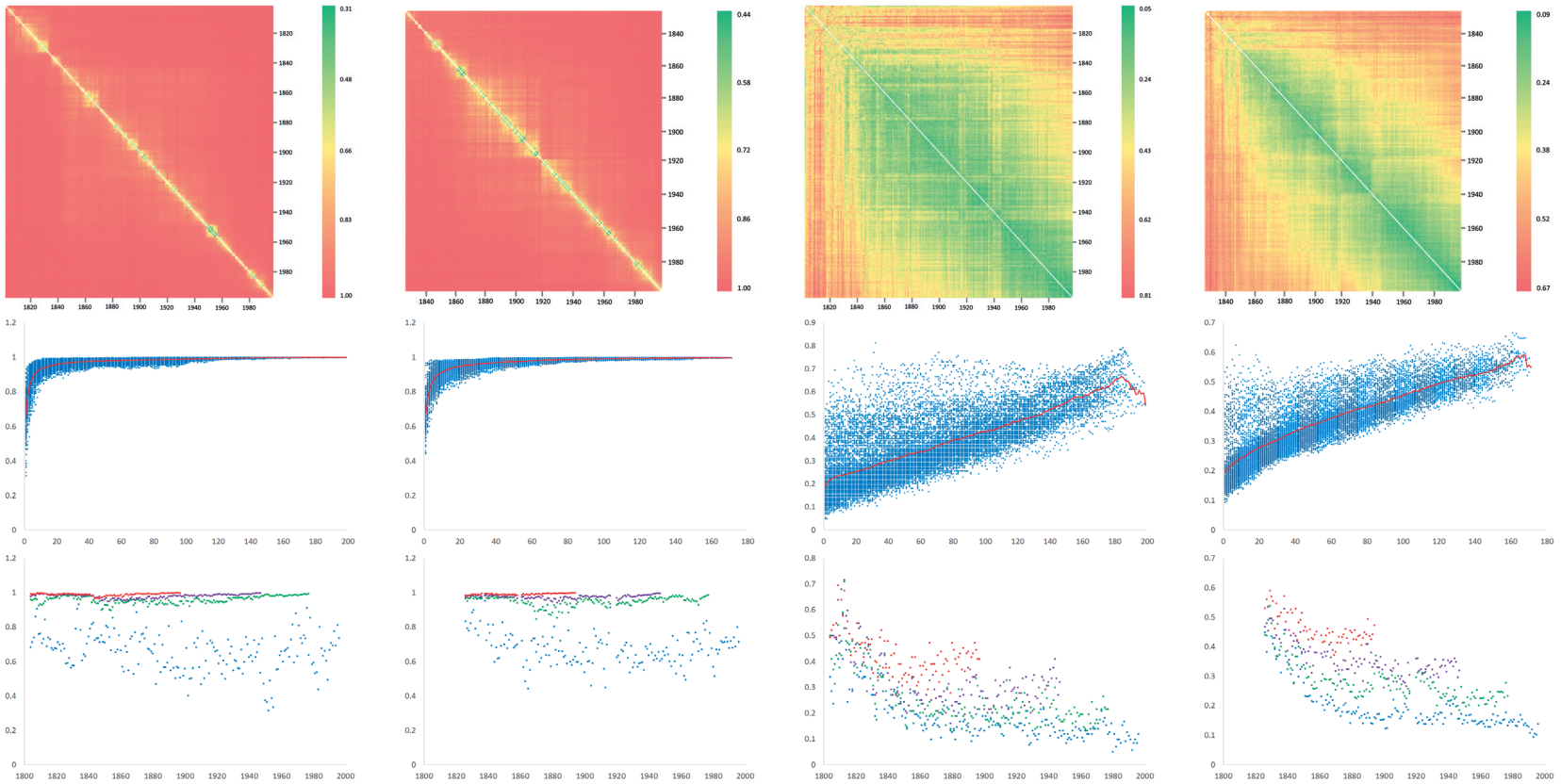


FIGURE 12.6 – **(1)** : Distances de Jaccard sur les 5-grammes de GDL ; **(2)** : Distances de Jaccard sur les 5-grammes de JDG ; **(3)** : Distances nucléaires sur les 5-grammes de GDL ; **(4)** : Distances nucléaires sur les 5-grammes de JDG ; **Haut** : Heatmap de la matrice des distances ; **Milieu** : Distances (bleu) et moyenne de ces distances (rouge) en fonction du nombre d'années de différence entre les sous-corpus ; **Bas** : Distances entre les années y_i et y_{i+n} avec $n = 1$ (bleu), $n = 20$ (vert), $n = 50$ (violet) and $n = 100$ (rouge)

Nous observons que la matrice des distances de Jaccard a un comportement global très similaire pour les niveaux 2 et 3 par rapport au niveau 1. Toutefois, en montant dans les niveaux n, la distance de Jaccard semble avoir des difficultés pour distinguer les différents sous-corpus. Cela se traduit par un tassement des distances vers la valeur maximale 1 et une plus grande variabilité des distances entre les années y_i et y_{i+1}. Ces effets sont particulièrement remarquables pour les niveaux $n = 4$ et $n = 5$ dont la distance de Jaccard ne donne que peu d'informations sur l'évolution linguistique du corpus.

Concernant la distance nucléaire, nous observons que son potentiel de différenciation est conservé au long des niveaux n. Toutefois, comme pour la distance de Jaccard, la distance nucléaire commence à perdre de sa précision (variabilité élevée) dès le niveau 4 (en particulier pour GDL). Nous ne pouvons donc pas tirer d'analyse robuste de ces résultats pour les niveaux supérieurs à 3. Toutefois, de façon intéressante et outre les effets de variabilité plus importants au niveaux supérieurs, les informations et visualisations des matrices distances nous montrent des courbes dont le comportement a tendance à être les mêmes quel que soit le niveau n. Nous observons que la distance nucléaire semble mesurer une vitesse de changement linguistique des deux corpus qui devient constante avec les années. Les périodes de guerre semble affecter l'évolution des distances sur le court terme uniquement. La distance nucléaire semble filtrer une partie importante des variabilités observées dans les années récentes, objet de plusieurs perturbations de reconnaissance OCR et de bruit. Globalement, la méthode de la distance nucléaire montre une plus grande stabilité avec le niveau n que la distance de Jaccard.

Lors de l'analyse du niveau $n = 1$, des simulations ont été effectuées sur la base d'une distribution zipfienne et de la taille réelle du corpus comme un échantillon tiré de cette distribution. Nous en avions conclu que la distance nucléaire était aussi impactée par l'évolution de la taille du corpus malgré la réduction constatée de cet effet. Toutefois, en raison de la stabilité exceptionnelle de la distance nucléaire vis-à-vis des différentes hypothèses concernant la taille du vocabulaire, nous avons proposé de calculer la différence entre les distances réelles calculées sur le corpus et les distances simulées ne contenant aucun effet d'évolution linguistiques. Nous en avons conclu que la distance résultante évolue de façon constante. Les mêmes résultats sont obtenus à des niveaux plus élevés de n et le comportement macroscopique des distances ne semblent pas montrer d'évolution particulière sur ces corpus.

Nous ne pouvons donc pas conclure à une quelconque accélération ou décélération de l'évolution linguistique, mais plutôt que la langue au travers de ce corpus de presse semble évoluer de façon constante sans être durablement influencée par diverses perturbations du corpus comme par exemple les événements historiques importants que sont les deux guerres mondiales qui n'ont perturbé l'évolution linguistique du corpus de presse que localement. Ces distances sont basées sur la présence ou absence des n-grammes pour la distance de Jaccard et sur l'ordre fréquentiel de ces n-grammes pour la distance nucléaire. Dans la prochaine section, nous observons l'évolution des n-grammes au travers de leur distribution fréquentielle annuelle via la notion d'entropie qui permet, au contraire de la mesure des distances, de décomposer la contribution spécifique de chaque n-gramme à la mesure.

12.2 Entropie

Dans cette section nous reprenons la notion d'entropie appliquée à chaque niveau n des n-grammes afin de déterminer l'évolution de cette mesure au cours du temps et sa relation avec le niveau n. Les calculs d'entropie diachronique pour chaque niveau n sont présentés dans les Figure 12.7 pour GDL et 12.8 pour JDG.

Nous y observons une augmentation monotone croissante avec des variations lors d'années particulières comme les années bruitées par la qualité des journaux scannés plus faible ou l'année 1998 ne contenant que deux mois. Toutefois, il est clair que l'entropie augmente avec la taille du corpus, rendant l'évolution diachronique difficile à interpréter.

Cependant, nous pouvons observer un phénomène intéressant via l'évolution de l'entropie annuelle moyenne en fonction du niveau n. Ces valeurs sont présentées dans la Figure 12.9. A l'instar du nombre d'éléments d'occurence 1 vue dans la section précédente, nous observons une asymptote de l'entropie en fonction du niveau n, ce qui signifie que la mesure de l'information supplémentaire apportée par le passage à un niveau n plus grand diminue au fur et à mesure que n augmente.

La Figure 12.10 présente la même information, mais permet la comparaison entre les deux journaux. Nous observons que l'entropie des deux journaux en fonction de n forme deux asymptotes différentes. Toutefois, la normalisation en terme de proportion supplémentaire d'informations par rapport au niveau précédent montre des valeurs identiques. Par exemple, la passage au niveau des 4-grammes n'apporte donc qu'environ 5% d'informations supplémentaires par rapport au niveau des 3-grammes. Ces dernières valeurs sont reprises dans la Table récapitulative 12.1.

	GDL			JDG		
	Entropie	**Différence**	**Proportion**	**Entropie**	**Différence**	**Proportion**
1	7.48			7.63		
2	11.98	4.50	60.2%	12.26	4.62	60.6%
3	14.01	2.04	17.0%	14.39	2.13	17.4%
4	14.67	0.66	4.7%	15.10	0.72	5.0%
5	14.87	0.20	1.4%	15.33	0.23	1.5%
6	14.95	0.07	0.5%	15.42	0.09	0.6%
7	14.98	0.03	0.2%	15.47	0.04	0.3%
8	15.00	0.02	0.1%	15.49	0.03	0.2%
9	15.01	0.01	0.1%	15.51	0.02	0.1%

TABLE 12.1 – Entropie moyenne en fonction du niveau n avec différences et proportions

Si l'analyses des distances de la section précédente nous a conduit à considérer uniquement les niveaux $n > 4$, l'analyse de l'entropie confirme que peu d'informations globales supplémentaires sont disponibles aux niveaux supérieurs n en terme d'analyse à l'échelle Macro.

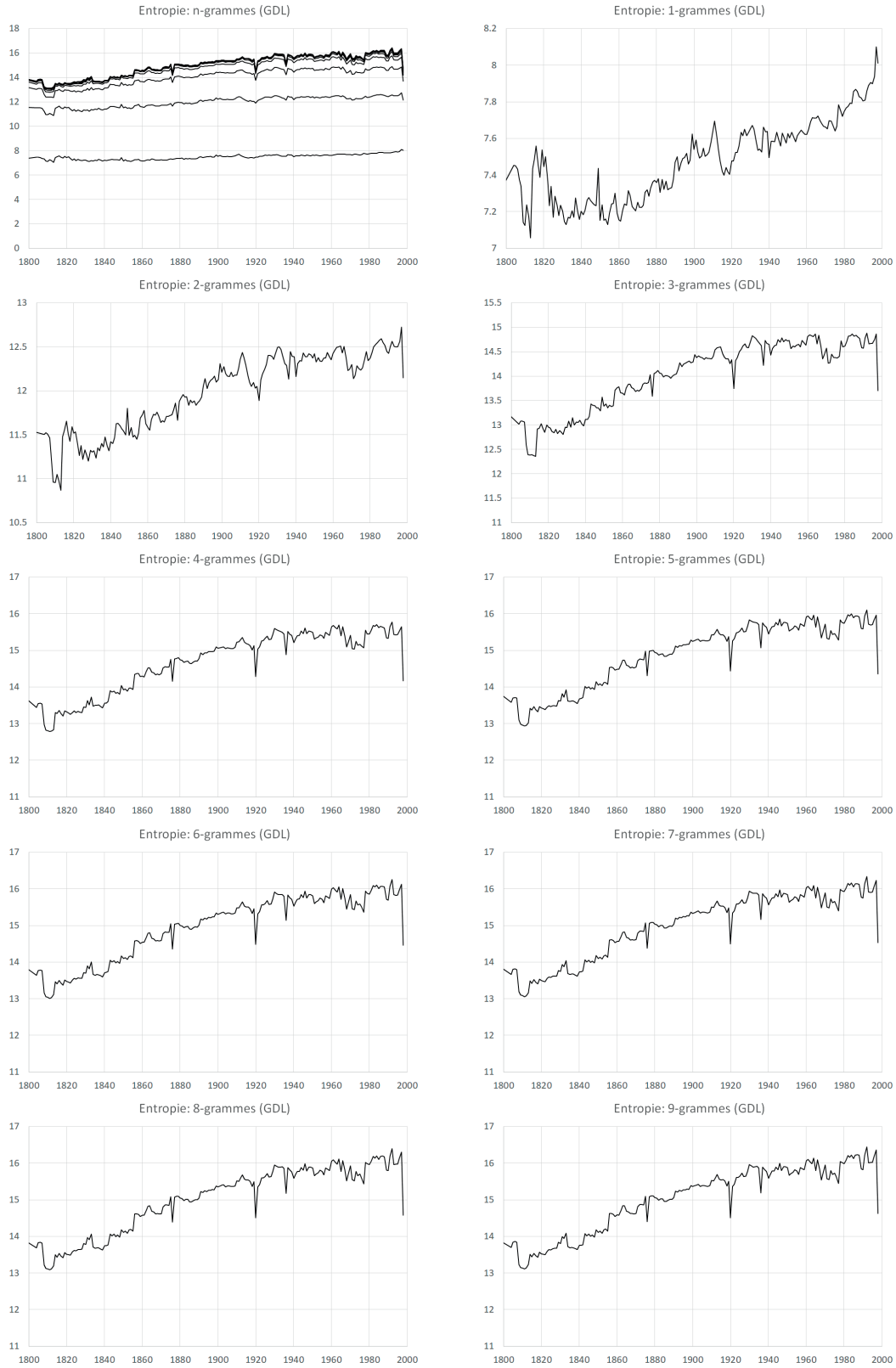


FIGURE 12.7 – Entropie des sous-corpus annuels de GDL

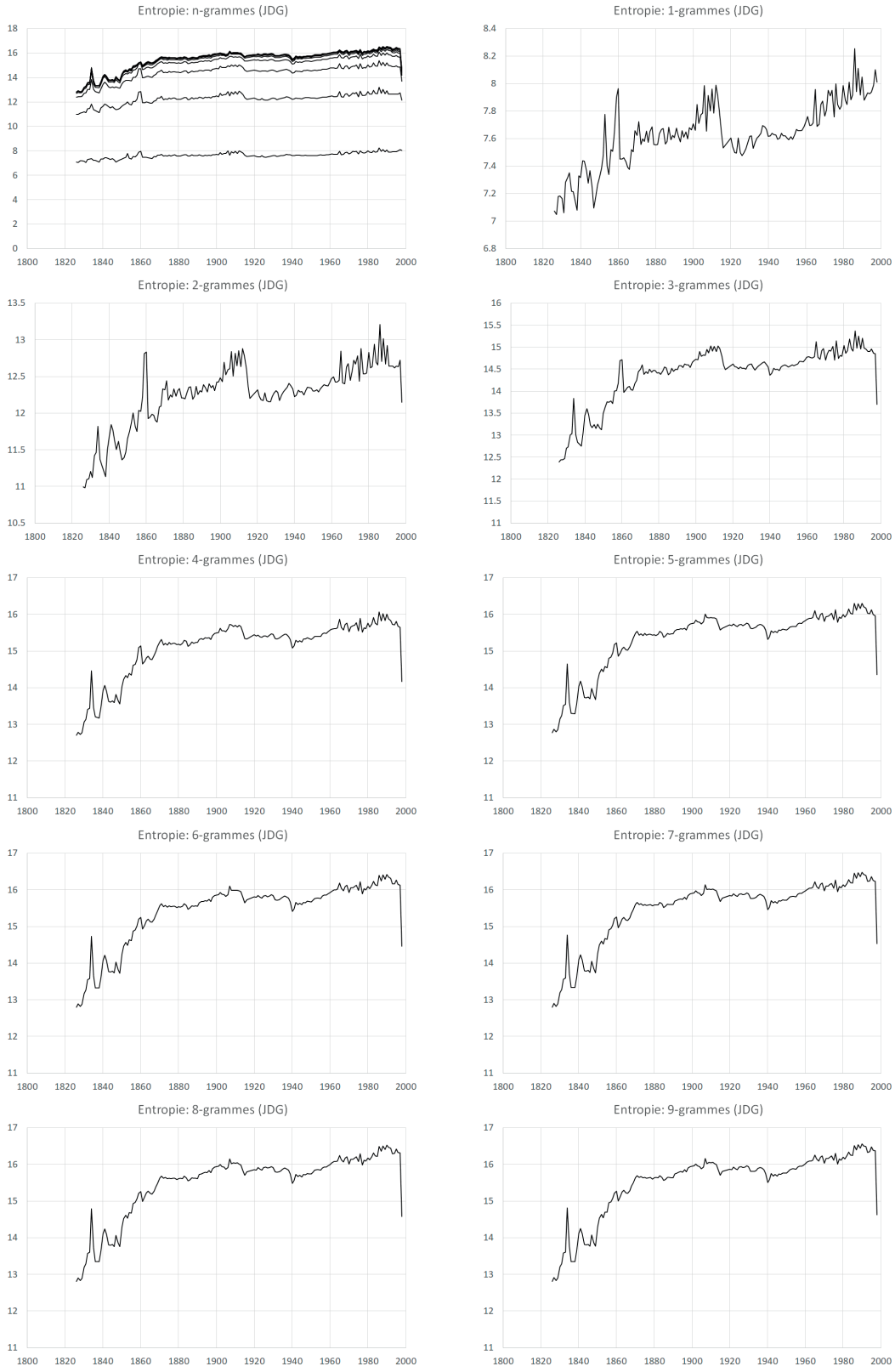


FIGURE 12.8 – Entropie des sous-corpus annuels de JDG

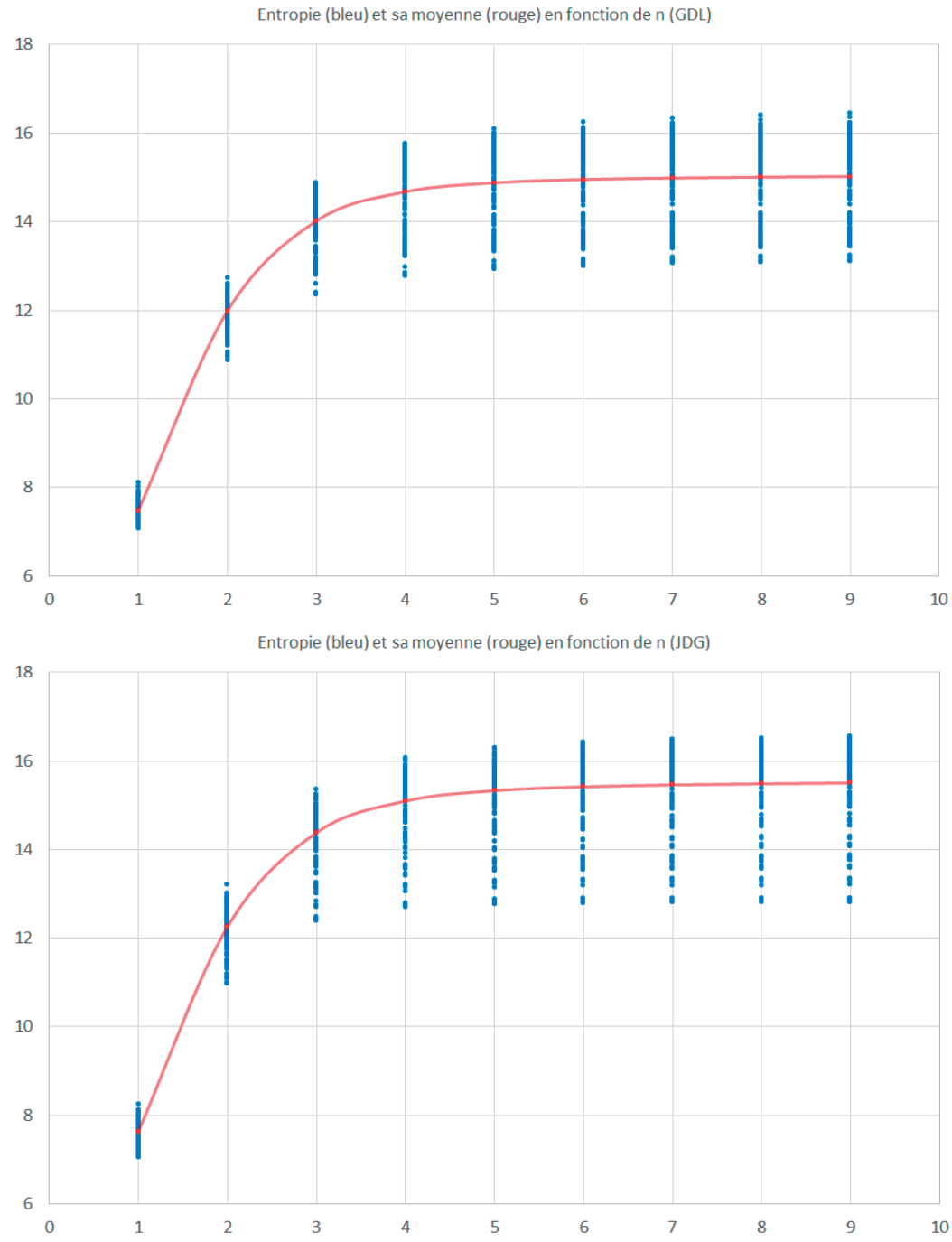


FIGURE 12.9 – Entropie annuelle et entropie annuelle moyenne en fonction du niveau n pour GDL (haut) et JDG (bas)

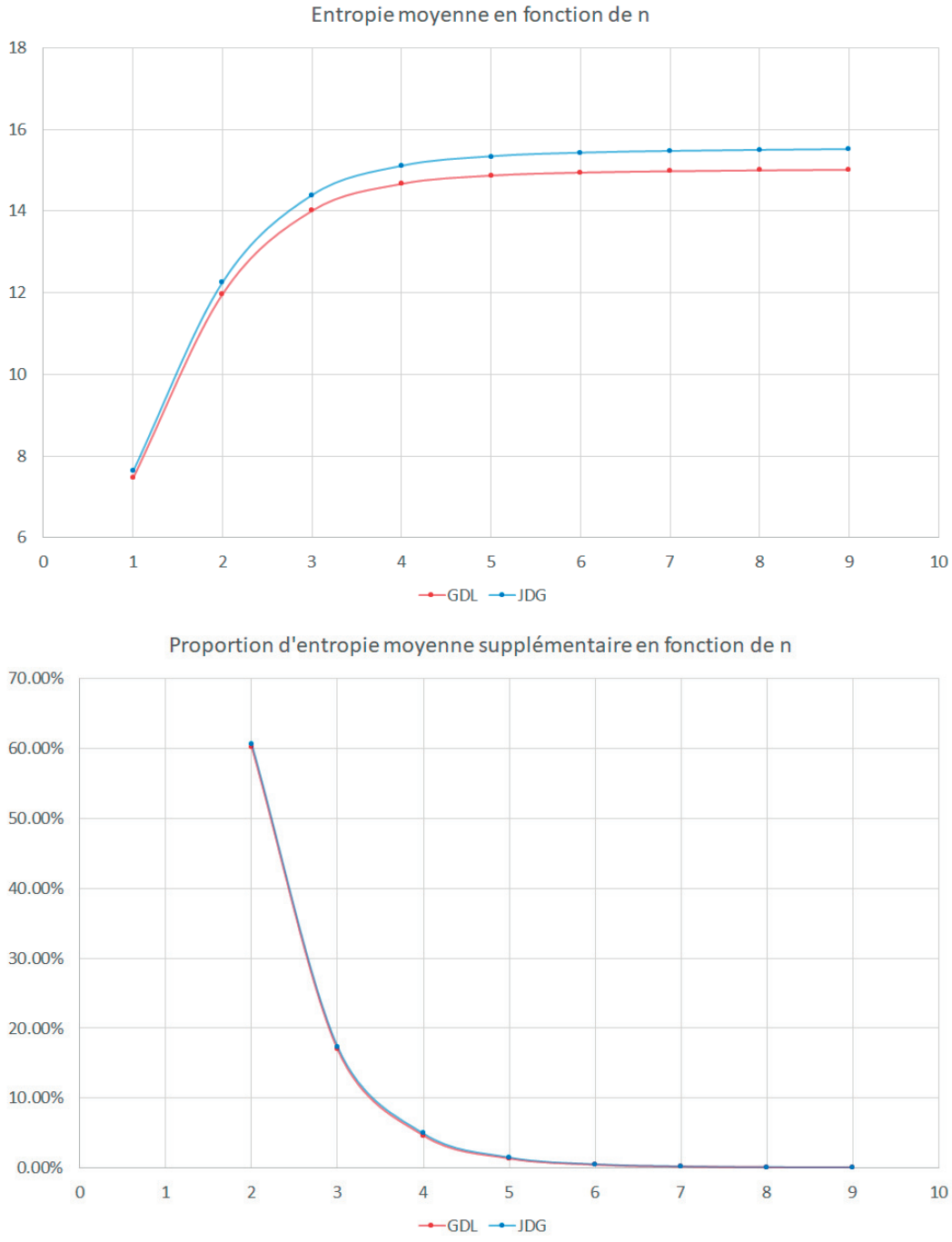


FIGURE 12.10 – Entropie annuelle moyenne en fonction du niveau n pour GDL et JDG en valeurs absolues (haut) et en valeurs relatives (bas)

Nous pouvons utiliser la même idée de combiner la mesure de l'entropie avec la notion de noyau résilient en réduisant simplement la mesure totale à la contribution de chaque n-gramme du noyau résilient. Afin de permettre la comparaison entre les journaux, nous utiliserons l'intersection des noyaux résilient de chaque journal comme ensemble de référence.

La Figure 12.11 présente la mesure de l'entropie que le noyau résilient commun à JDG et GDL et ce pour chaque niveau n de 1 à 5 et chaque journaux.

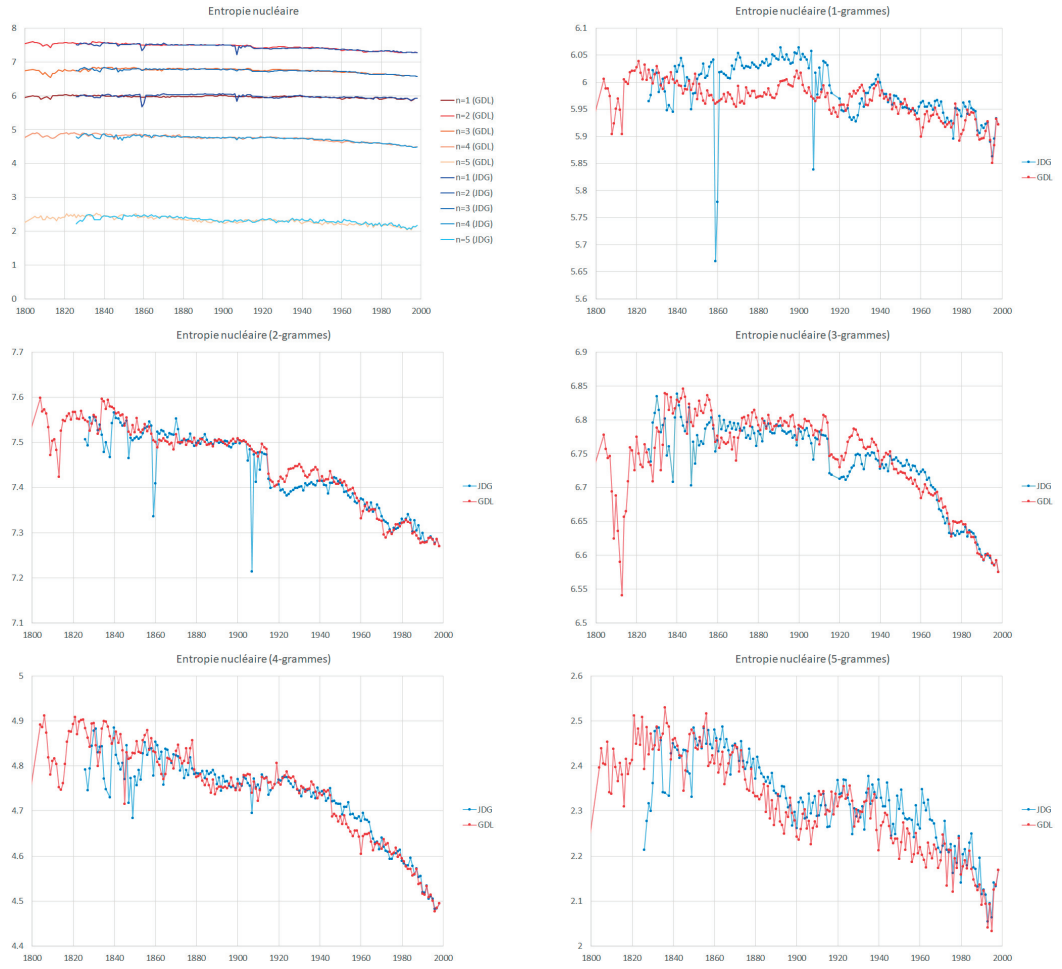


FIGURE 12.11 – Entropie du noyau résilient commun en fonction des années pour les corpus de GDL et JDG

Nous observons pour le niveau $n = 2$ et $n = 3$, les trois valeurs extrêmes correspondant aux années perturbées par des erreurs d'OCR récurrentes sur le corpus de JDG. La mesure d'entropie nucléaire permet donc au minimum de déceler ces perturbations dans le corpus.

Nous observons, pour le niveau $n = 2$ uniquement, une période de 1850 à 1920 ou le noyau montre une entropie supérieure du corpus de JDG par rapport à GDL. Cet effet s'estompe ensuite dans les niveau n supérieurs.

Enfin, nous observons pour chaque niveau et pour les deux journaux une tendance persistante de diminution de l'entropie au cours du temps, en particulier à partir de 1930. Cela suggère une évolution de la forme de la distribution fréquentielle des n-grammes du noyau résilient (c.f. Figure 12.12).

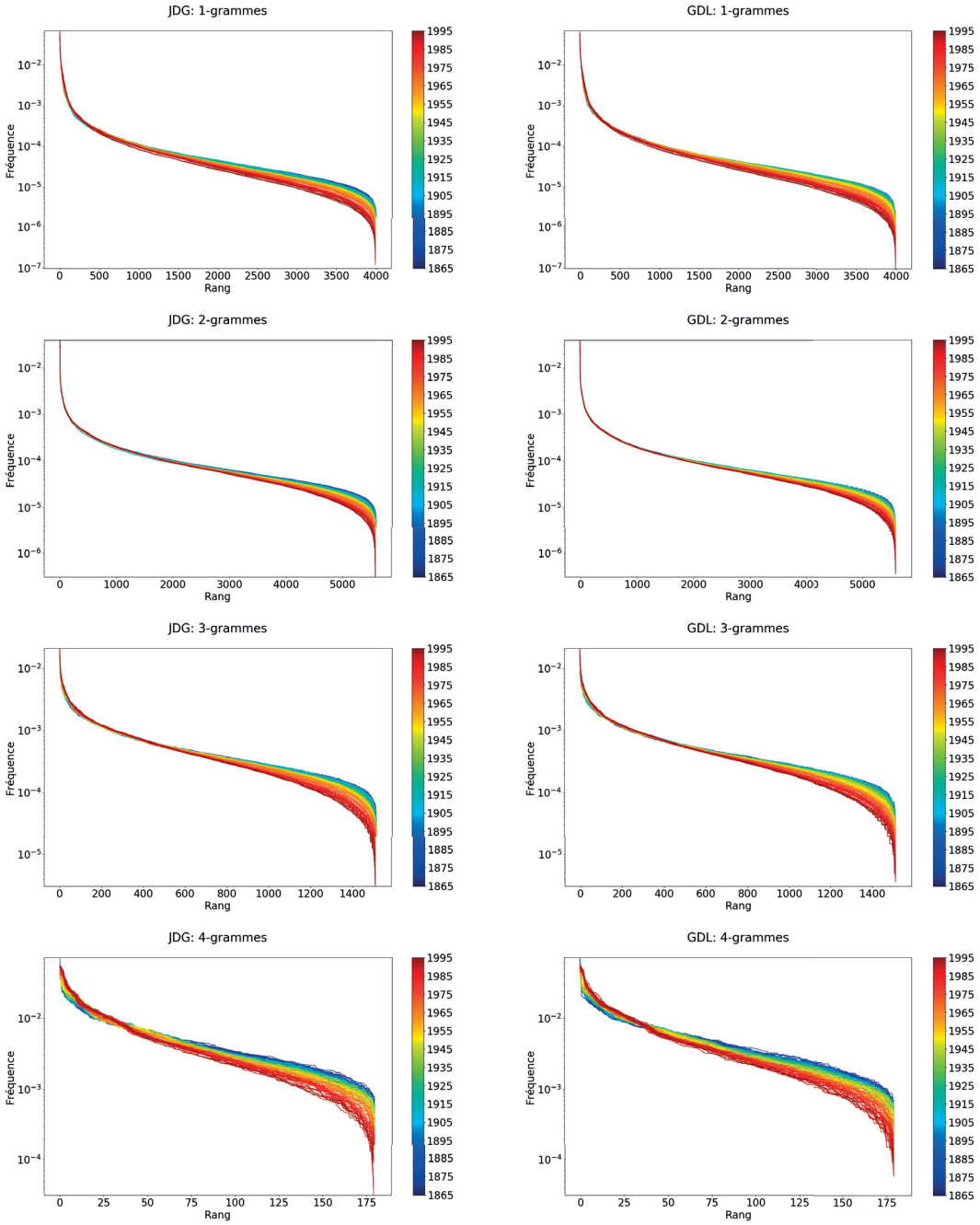


FIGURE 12.12 – Distribution fréquentielle des n-grammes du noyaux résilient avec n=1, 2, 3 et 4 pour JDG (gauche) et GDL (droite) de 1865 à 1995

Nous avons également calculé un coefficient de régression linéaire sur les différentes contributions à l'entropie des n-grammes composant les noyaux résilients annuels dans la période de 1930 à 1997 pour chaque niveau de n et pour les prétraitements alpha et alphanumérique des deux journaux.

Ce faisant, nous exploitons la possibilité de regarder cette mesure diachronique selon sa composition pour chaque n-gramme. Cela nous permet d'observer à la loupe les contributions de ces n-grammes et en particulier celles qui ont particulièrement baissé ou augmenté dans la période allant de 1930 à 1997. Cependant, il est clair que cette façon de faire n'explique qu'en partie les changements importants de l'entropie nucléaire car ceux-ci devraient être regardés comme un effet global résultant de l'évolution de la forme de la courbe de distribution fréquentielle du noyau résilient.

Ainsi, pour le niveau $n = 1$, nous observons que les mots dont la contribution à l'entropie globale est en forte diminution sont des mots aussi fréquents que "et", "la", "que", "de", "il", "à", "les", "on", etc. En effet, ces éléments fréquents ont tendance à baisser de fréquence dans les années les plus récentes, comme illustrée dans les Figures 12.13 et 12.14.

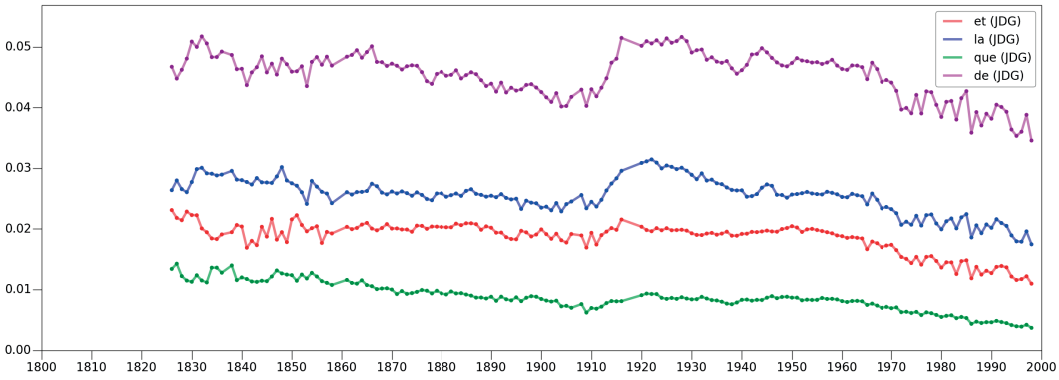


FIGURE 12.13 – Profil fréquentiel des mots "et", "la", "que" et "de" dans JDG

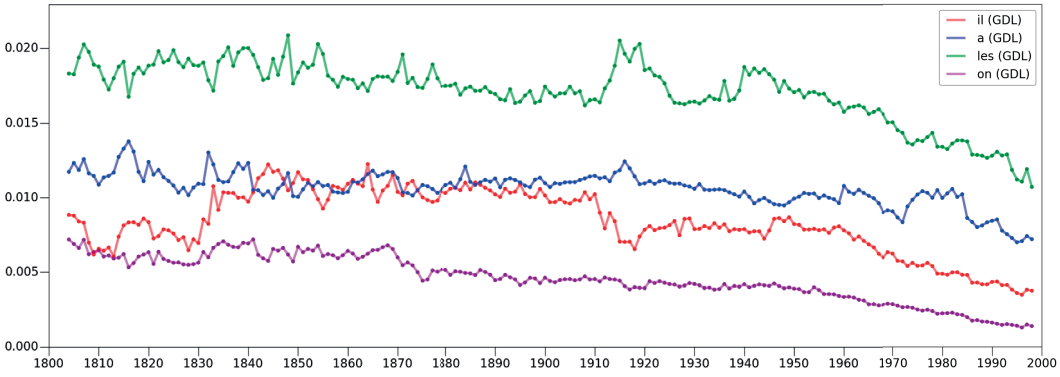


FIGURE 12.14 – Profil fréquentiel des mots "il", "à", "les" et "on" dans GDL

D'autres mots ont une contribution à l'entropie globale en forte augmentation et viennent contrebalancer la baisse d'entropie. Par exemple les mots "selon", "face", "centre", "développement", "également", "notamment", "monde", "Europe" ont tendance à augmenter de fréquence dans les années les plus récentes, comme illustré dans les Figures 12.15 et 12.16.

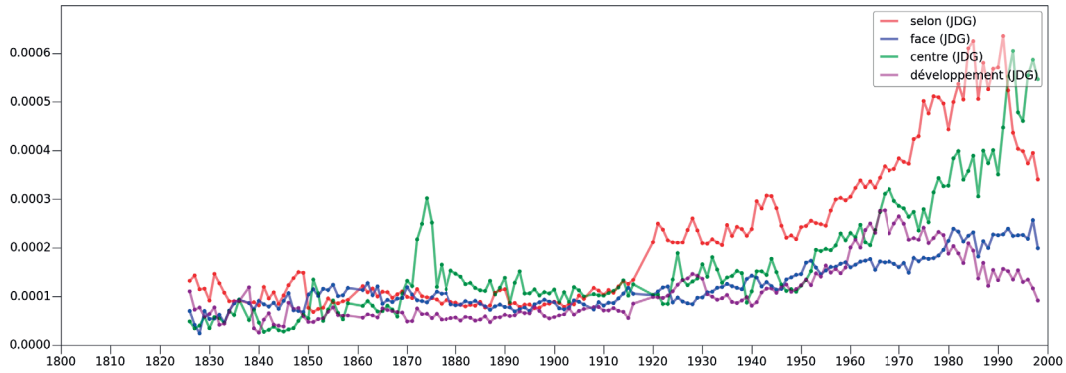


FIGURE 12.15 – Profil fréquentiel de "selon", "face", "centre" et "développement" dans JDG

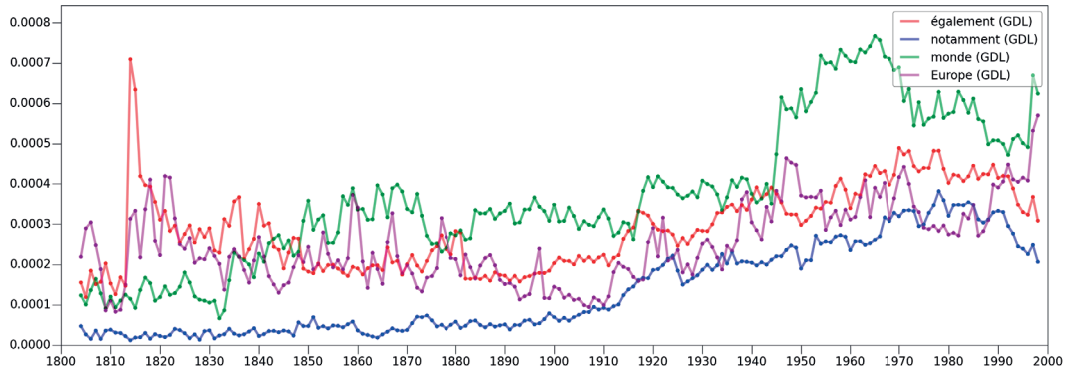


FIGURE 12.16 – Profil fréquentiel de "également", "notamment", "monde" et "Europe" dans GDL

Bien entendu, l'évolution décroissante de la fréquence des mots parmi les plus fréquents est également la conséquence de l'augmentation fréquentielle des chiffres qui prennent de plus en plus de place dans le corpus vis-à-vis des autres mots qui ne sont pas des données numériques. Toutefois, nous avons observé globalement la même tendance si l'on considère le prétraitement alpha plutôt que alphanumérique.

Si les nouvelles sections de journaux peuvent jouer un rôle dans l'augmentation de fréquence de certain mots (comme par exemple les lettres de l'alphabet seules ou les abréviations), d'autres mots tels que "selon", "face", "également" ou "notamment" ont une connotation plus neutre et peuvent être le résultat d'un effet d'évolution de langage.

De plus, les effets de diminution de l'entropie sont constatés à tous les niveaux n et nous pouvons illustrer d'autres contributions positives et négatives de n-grammes à cette mesure d'entropie globale du noyau résilient.

Les Figures 12.17 et 12.18 présentent le profil fréquentiel de n-grammes ayant une contribution à l'entropie qui diminue dans la période des années les plus récentes. Il s'agit des n-grammes "qu'il", "qu'on", "point de vue", "un grand nombre de", "il y a lieu", "à l'égard de", "ce qu'il y a" et "de la manière la plus".

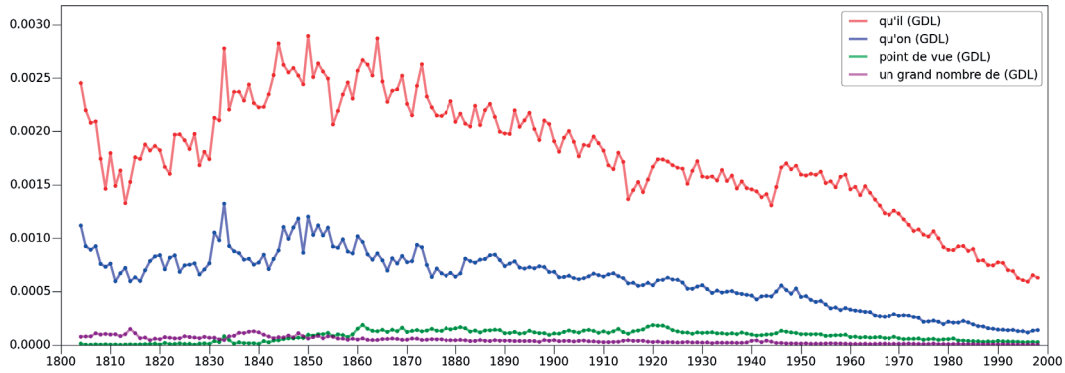


FIGURE 12.17 – Profil fréquentiel de "qu'il", "qu'on", "point de vue" et "un grand nombre de" dans GDL

Les n-grammes "qu'il", "qu'on", "point de vue" et "un grand nombre de" font partie du noyau résilient (selon le niveau n) et sont fréquents. En outre, ce sont de expressions qui s'utilisent dans un grand nombre de contextes différents et elles ne sont donc pas rattachées à des situations spécifiques. Pourtant, il est surprenant de constater que la fréquence de ces n-grammes affichent une tendance claire à diminuer avec le temps et en particulier dans les 30 dernières années du corpus.

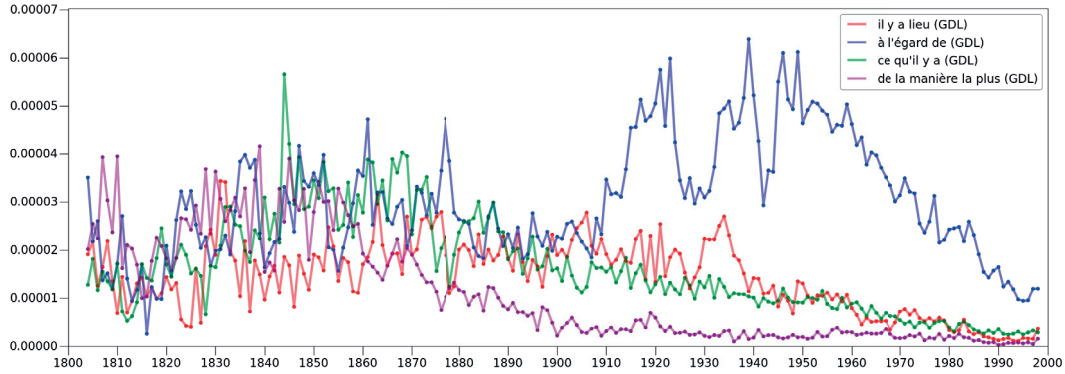


FIGURE 12.18 – Profil fréquentiel de "il y a lieu", "à l'égard de", "ce qu'il y a" et "de la manière la plus" dans GDL

Les n-grammes "il y a lieu", "à l'égard de", "ce qu'il y a" et "de la manière la plus" diminuent également dans les années les plus récentes bien que certains comme "à l'égard de" semblent avoir un régime de colline ayant augmenté fortement au début du XXe siècle pour diminuer comme les autres n-grammes dans les 50 dernières années du corpus. De la même façon, ces n-grammes peuvent être utilisés dans de nombreux contextes et ils ne sont donc pas directement rattachés à des situations spécifiques ou à des événements particuliers.

Toutefois, l'utilisation de ces expressions sont en nette baisse dans les corpus de JDG et GDL. Il est probable que le corpus de presse reflète par l'évolution des formulations utilisées et notamment celles utilisées dans de nombreux contextes différents, des évolutions linguistiques plus globale dans la langue.

D'autres n-grammes dont la contribution à l'entropie diminue avec le temps sont observés comme "alors que", "il s'agit", "lors de", "au sein de", "de plus en plus", "la plupart des", "de manière" ou "le nombre de" (cf. les Figures 12.19 et 12.20).

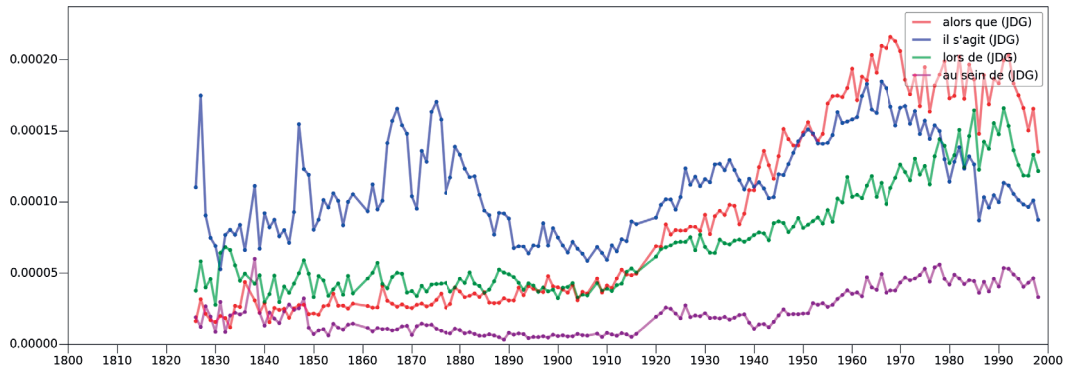


FIGURE 12.19 – Profil fréquentiel de "alors que", "il s'agit", "lors de" et "au sein de" dans JDG

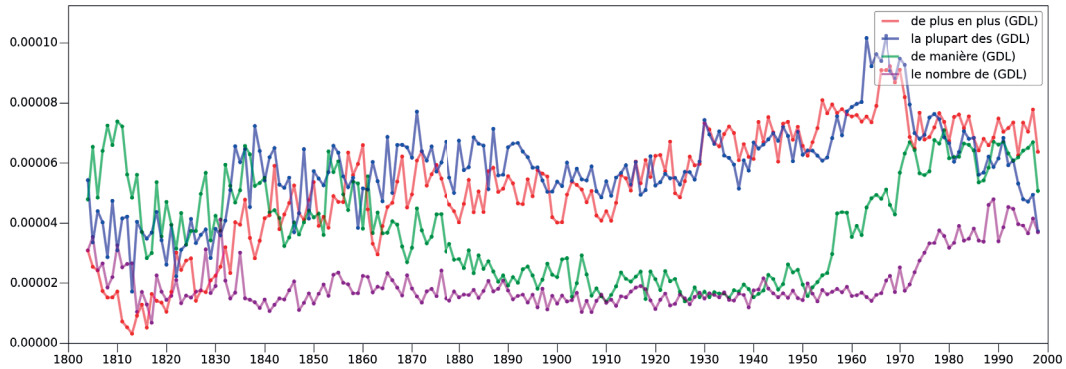


FIGURE 12.20 – Profil fréquentiel de "de plus en plus", "la plupart des", "de manière" et "le nombre de" dans GDL

Si certains n-grammes comme "alors que", "lors de", "au sein de" ou "le nombre de" ont une fréquence stable qui augmente ensuite brusquement dans les années les plus récentes, d'autres n-grammes subissent des variations importantes sur le long terme comme "il s'agit", "de plus ne plus", "la plupart des" ou "de manière.

Si l'on analyse le comportement de l'entropie nucléaire annuelle moyenne par rapport au niveau n (cf. Figure 12.21 et Table 12.2), nous avons une augmentation sur $n = 2$, pour avoir une baisse continue dès $n = 3$ dont l'entropie reste supérieure au niveau $n = 1$. Pour $n = 4$ et $n = 5$, l'entropie continue à décroitre de façon monotone. La mesure d'entropie nucléaire est similaire pour les deux journaux.

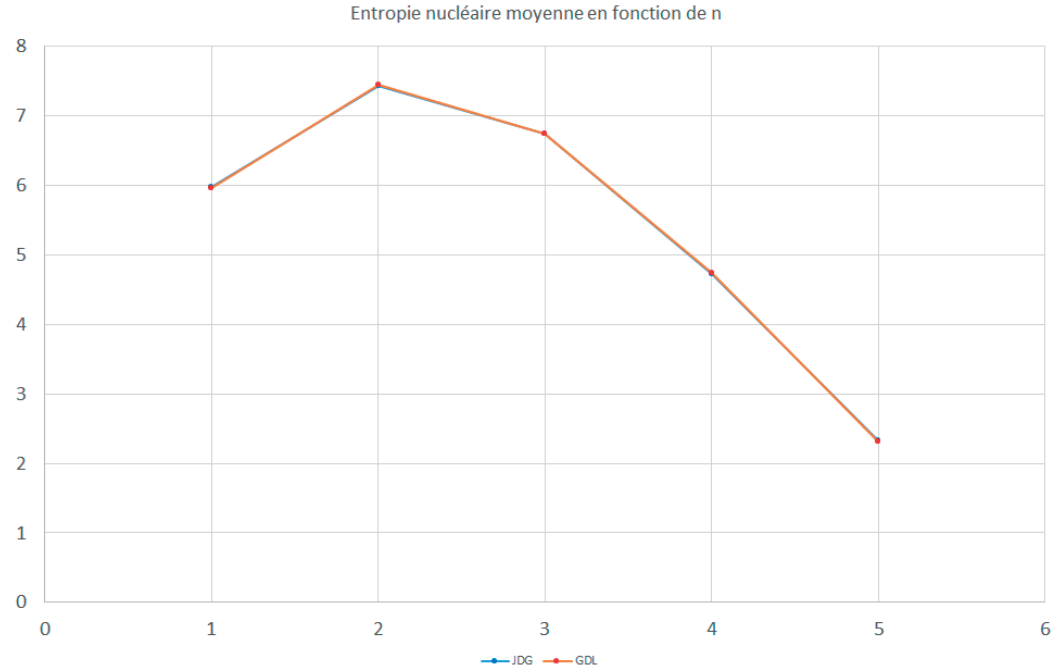


FIGURE 12.21 – Entropie nucléaire pour les niveau n de 1 à 5

	GDL			JDG		
	Entropie	**Différence**	**Proportion**	**Entropie**	**Différence**	**Proportion**
1	5.97			5.98		
2	7.46	1.49	24.9%	7.43	1.45	24.2%
3	6.74	-0.71	-9.6%	6.74	-0.70	-9.4%
4	4.75	-1.99	-29.6%	4.73	-2.01	-29.8%
5	2.32	-2.43	-51.2%	2.32	-2.41	-50.9%

TABLE 12.2 – Entropie nucléaire avec les différences absolues et relatives pour n allant de 1 à 5

Bien que corrélée avec la taille du noyau résilient qui varie également avec le niveau n, la mesure d'entropie suggère une quantité supérieure d'informations pour $n = 3$ par rapport à $n = 1$ alors que la taille du noyau résilient du niveau $n = 3$ est inférieure à $n = 1$.

La combinaison de la notion d'entropie et de noyau résilient permet non seulement la comparaison entre les deux corpus, mais également d'en mesurer l'évolution diachronique, révélant des informations globales sur l'évolution fréquentielle des éléments composant le noyaux résilient.

Le noyau résilient du niveau $n = 1$ comporte encore des erreurs d'OCR et des perturbations dues à l'apparition de sections particulières et répétitives du journal comme par exemple la bourse. En effet, nous y détectons notamment des variations importantes au niveau des mots formés d'une seule lettre et également d'un ou deux chiffres.

Toutefois, ce type d'erreur ou de perturbation est plus aléatoire que la formation classique des n-grammes munie d'une loi de type zipfienne et l'erreur dépend principalement de la longueur du n-grammes. En effet, si l'on considère l'exemple des mots formés d'une seule lettre et dans un cas aléatoire parfait, ce type de mot ne pourra s'identifier qu'à 26 possibilités différentes (sans compter les lettres particulières comme les lettres accentuées).

Avec seulement 26 possibilités, les mots formés de lettres seules sont présents chaque année et font donc partie du noyau résilient. Cependant à un niveau $n > 1$, l'espace des combinaisons de lettres seules augmente et le nombre de possibilités devient 26^n. Il est donc beaucoup plus rare de trouver ce type de combinaisons aléatoires dans le noyau résiliant à ces niveaux.

En résumé, nous avons observé une diminution importante de l'entropie sur les noyaux résilients de chaque niveau sur les années les plus récentes. En détaillant les contributions individuelles des différents n-grammes, nous avons observé que les n-grammes parmi les plus fréquents avaient tendance à diminuer, alors que d'autres n-grammes, parmi les moins fréquents, montaient en fréquence lors de ces mêmes années.

La somme de ces différentes contribution entrainent une diminution de l'entropie globale. Cet effet n'est pas dû à la variation de taille des corpus. En effet, nous avons effectué des simulations basées sur différentes lois de Zipf selon les tailles réalistes des ensembles de n-grammes et en considérant la taille réelle de nos données. Celles-ci nous ont permis de conclure que l'évolution de l'entropie nucléaire n'est quasiment pas impactée par les variations de taille des deux corpus JDG et GDL.

Une réserve existe toutefois concernant les premières années du corpus où les données, de taille trop faible, impliquent de trop grandes variations de fréquence et n'ont par conséquent que peu de représentativité générale du langage dont nous tentons de mesurer l'évolution.

La prochaine section mettra en avant les résultats des chronoclouds classiques, des chronoclouds différentiels et du visualisateur de n-grammes sur les niveaux $n > 1$ afin de mettre en évidence des évolutions fréquentielles de n-grammes particuliers.

12.3 Chronocloud

Chronocloud classique

Dans cette section nous présentons les chronoclouds pour chaque niveau $n > 1$ et nous détaillons ensuite nos observations en considérant l'ensemble des chronoclouds des différents niveaux, vu que nous considérons les niveaux n jusque $n = 9$, nous présentons 16 chronoclouds classiques dans les Figures 12.22 à 12.37.

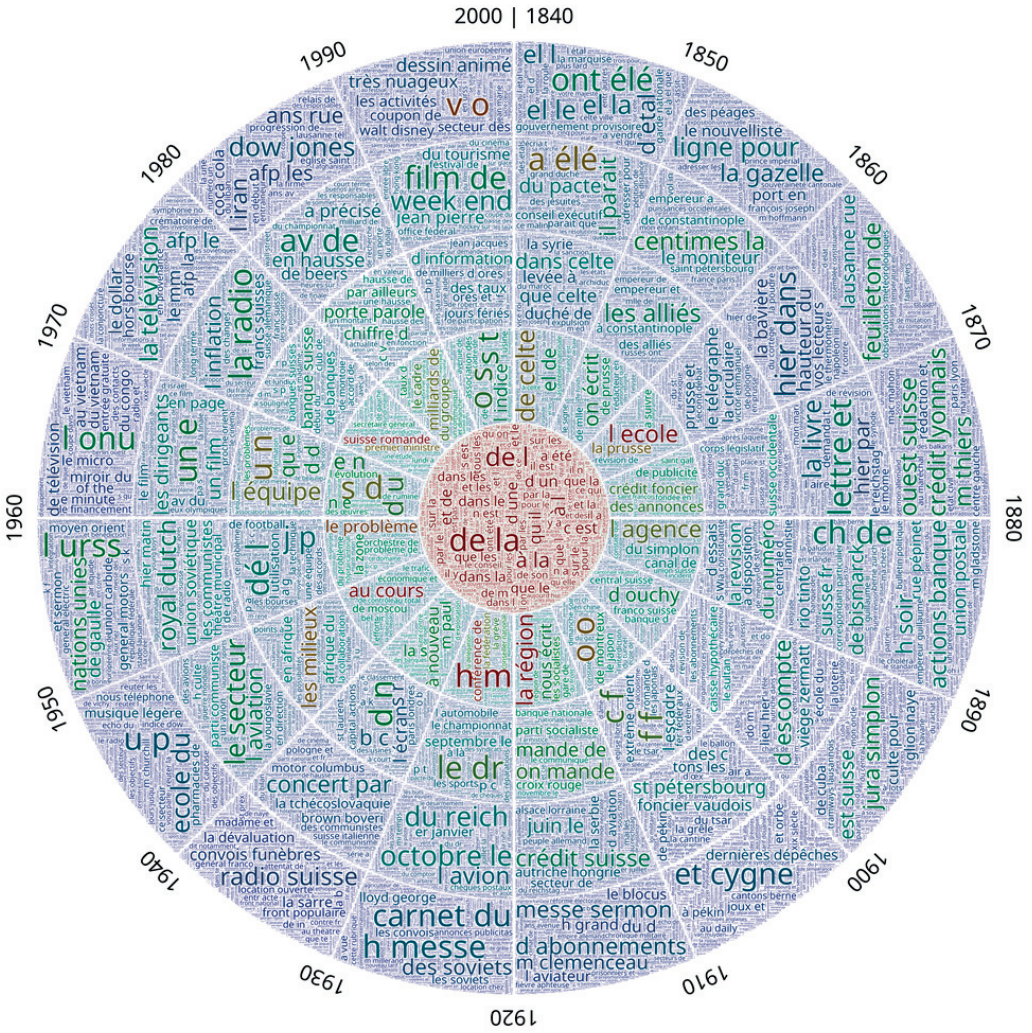


FIGURE 12.22 – Chronocloud classique sur les 2-grammes pour GDL

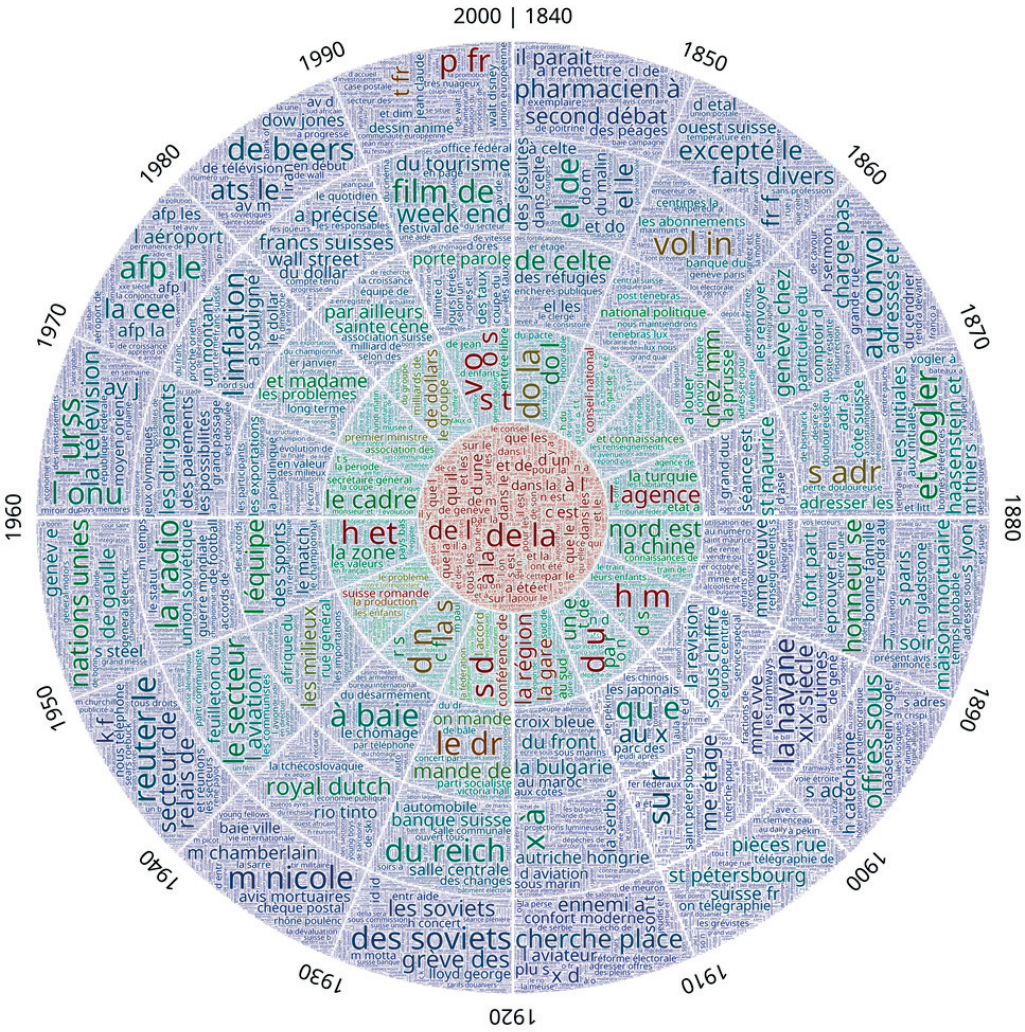


FIGURE 12.23 – Chronocloud classique sur les 2-grammes pour JDG

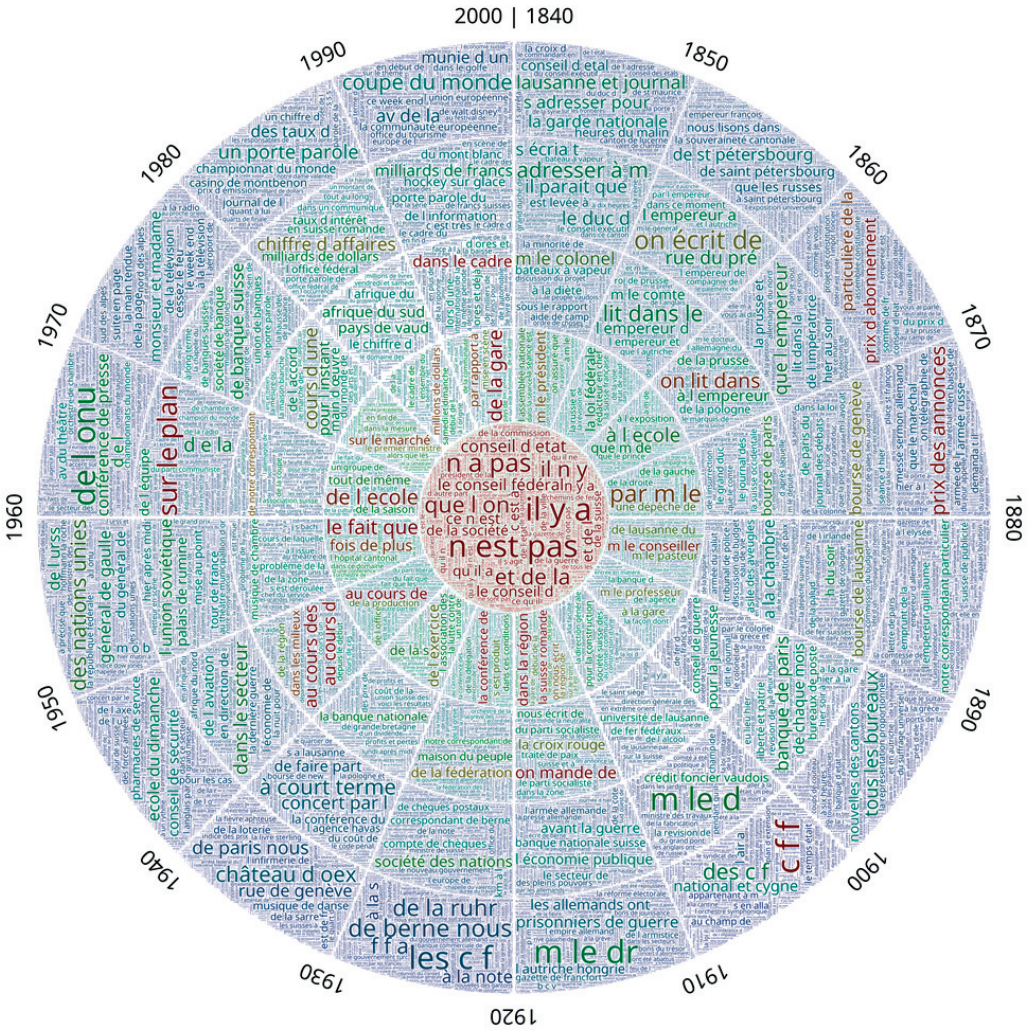


FIGURE 12.24 – Chronocloud classique sur les 3-grammes pour GDL

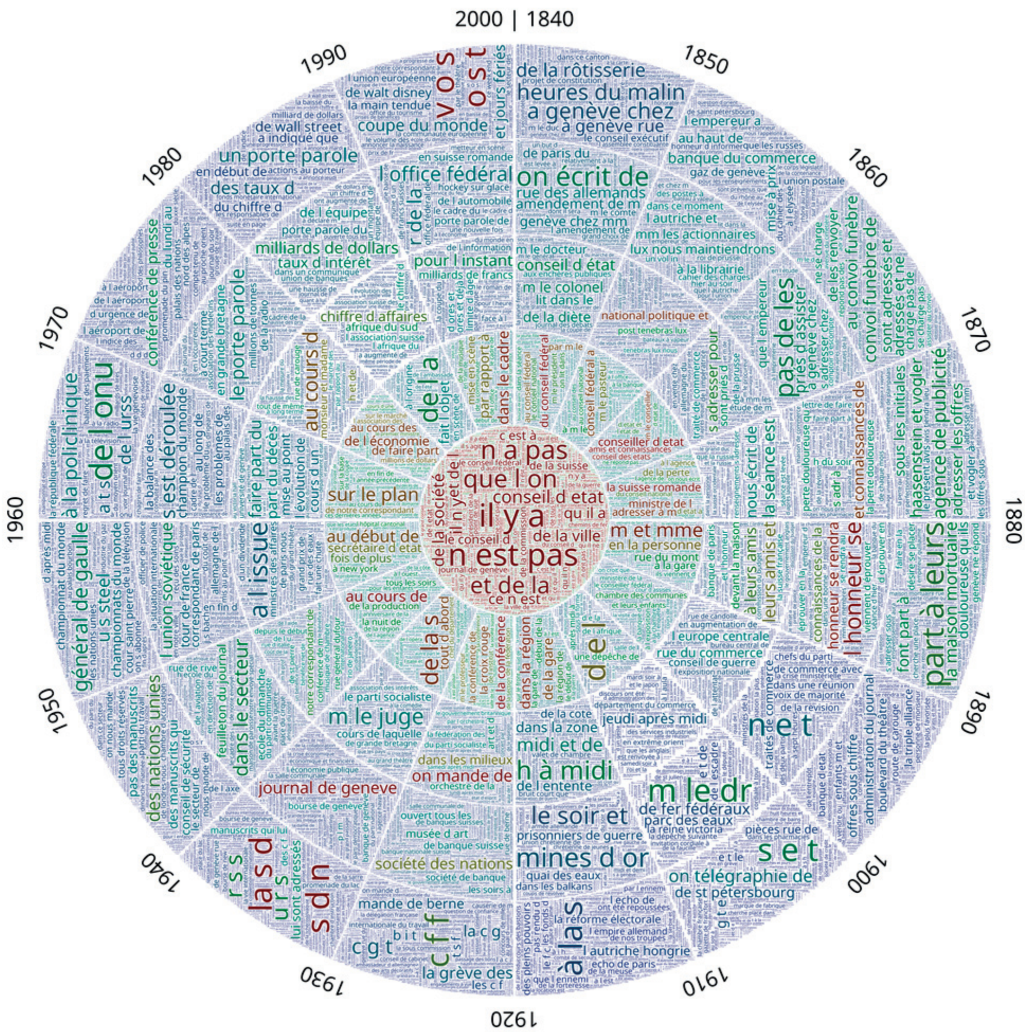


FIGURE 12.25 – Chronocloud classique sur les 3-grammes pour JDG

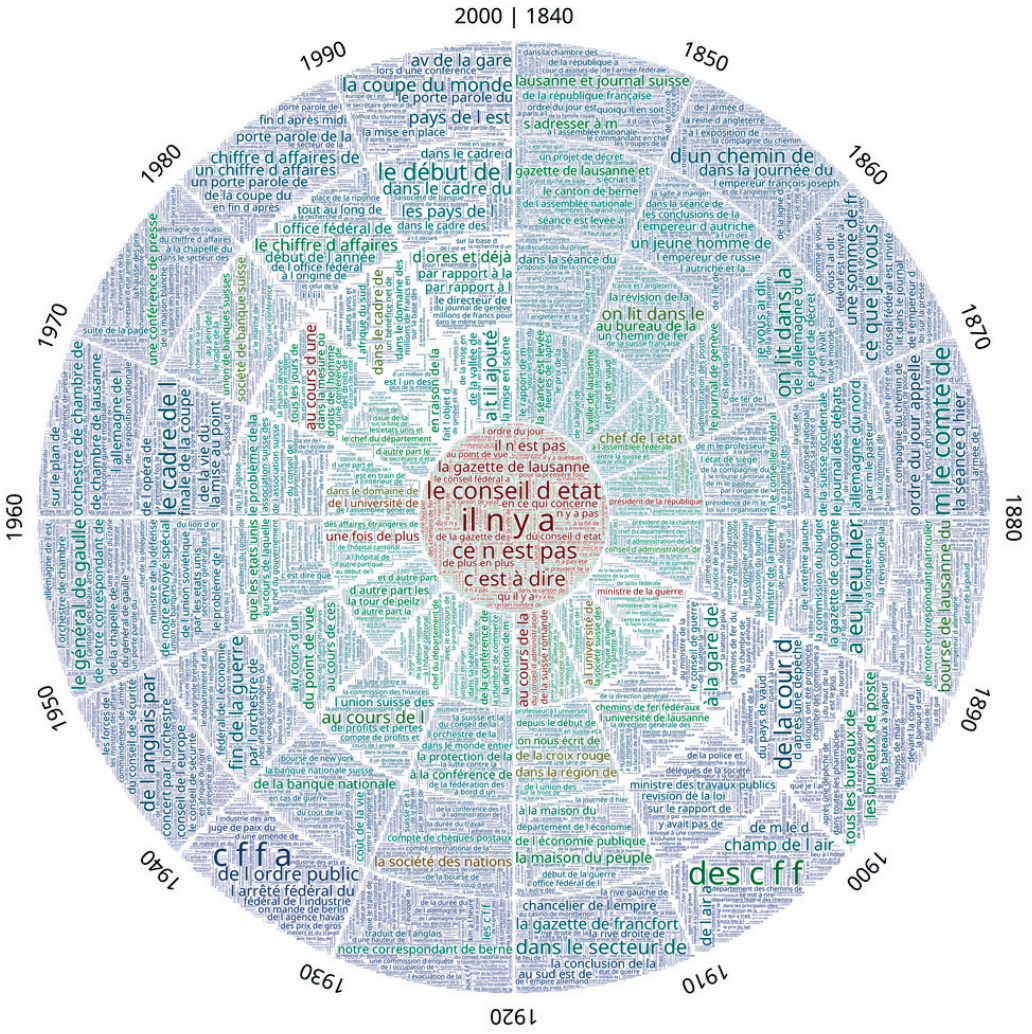


FIGURE 12.26 – Chronocloud classique sur les 4-grammes pour GDL

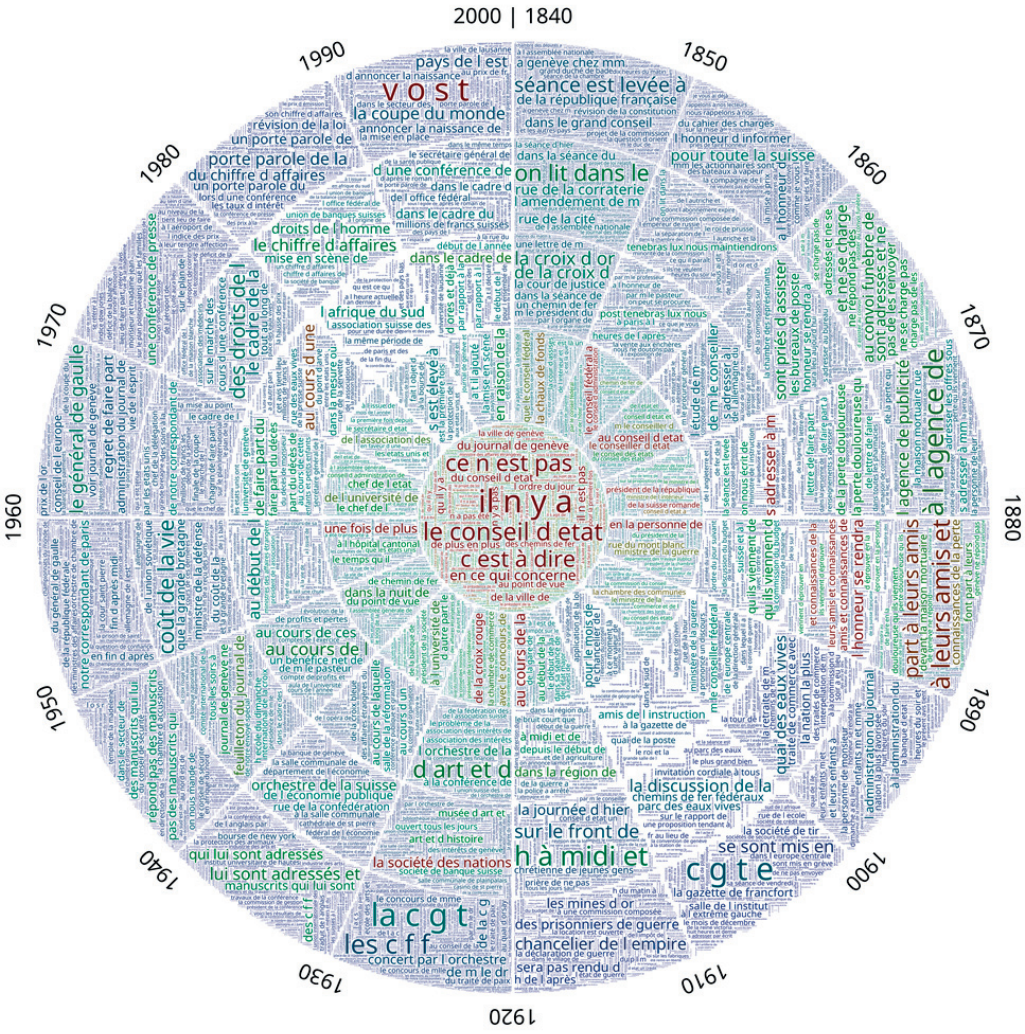


FIGURE 12.27 – Chronocloud classique sur les 4-grammes pour JDG

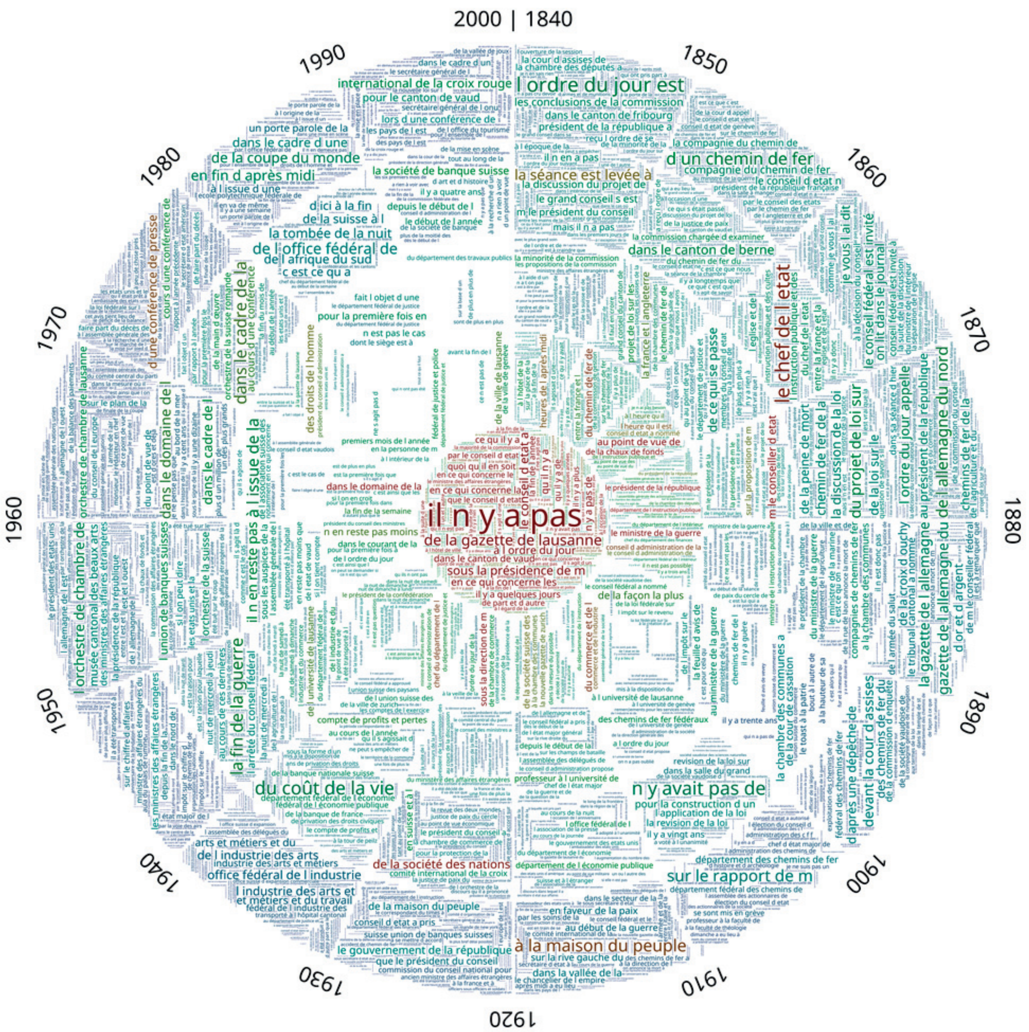


FIGURE 12.28 – Chronocloud classique sur les 5-grammes pour GDL

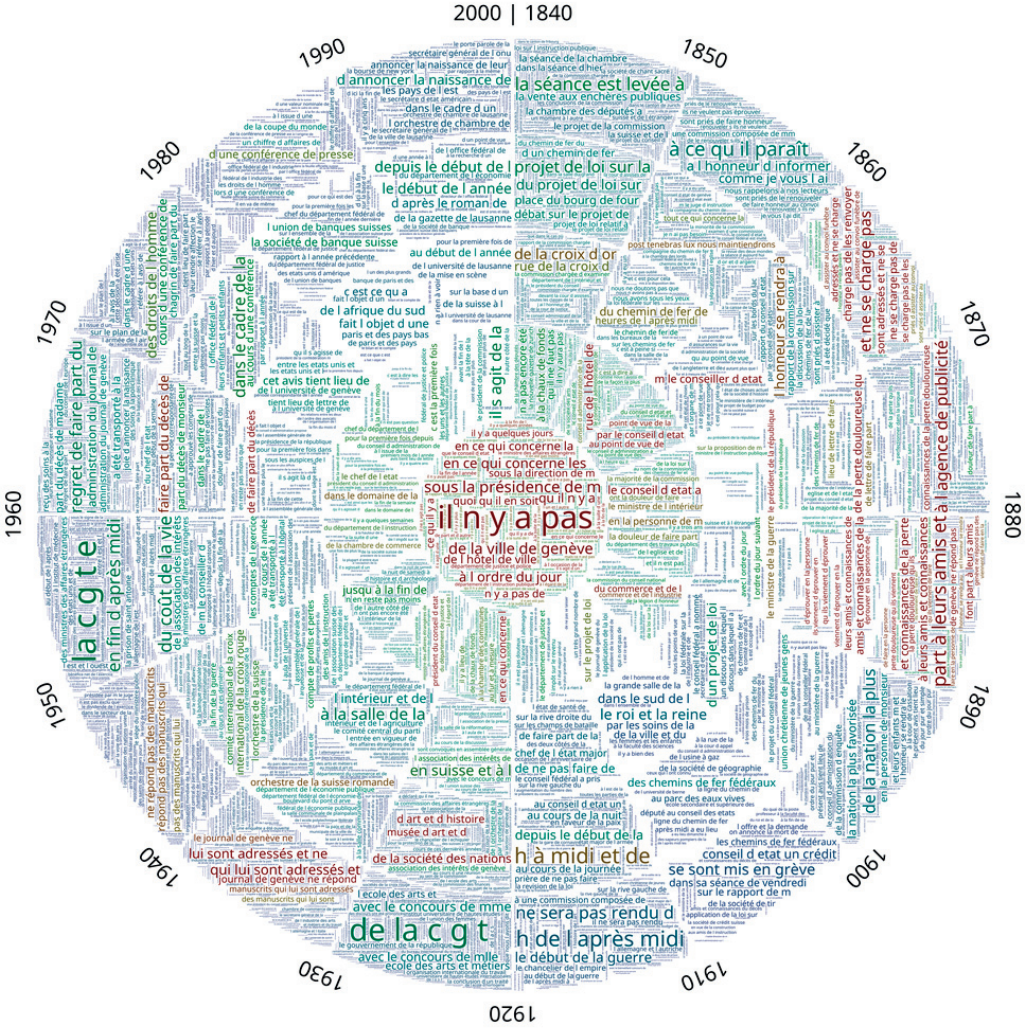


FIGURE 12.29 – Chronocloud classique sur les 5-grammes pour JDG

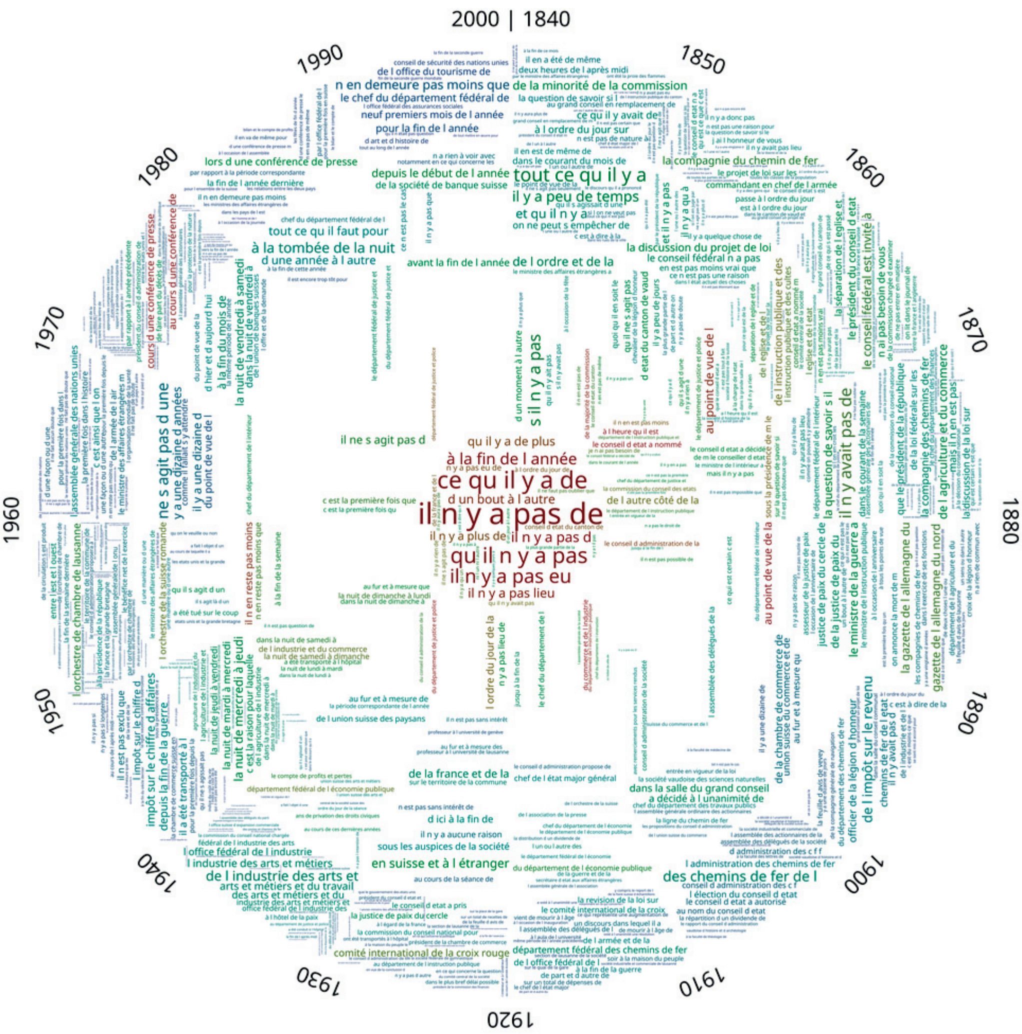


FIGURE 12.30 – Chronocloud classique sur les 6-grammes pour GDL

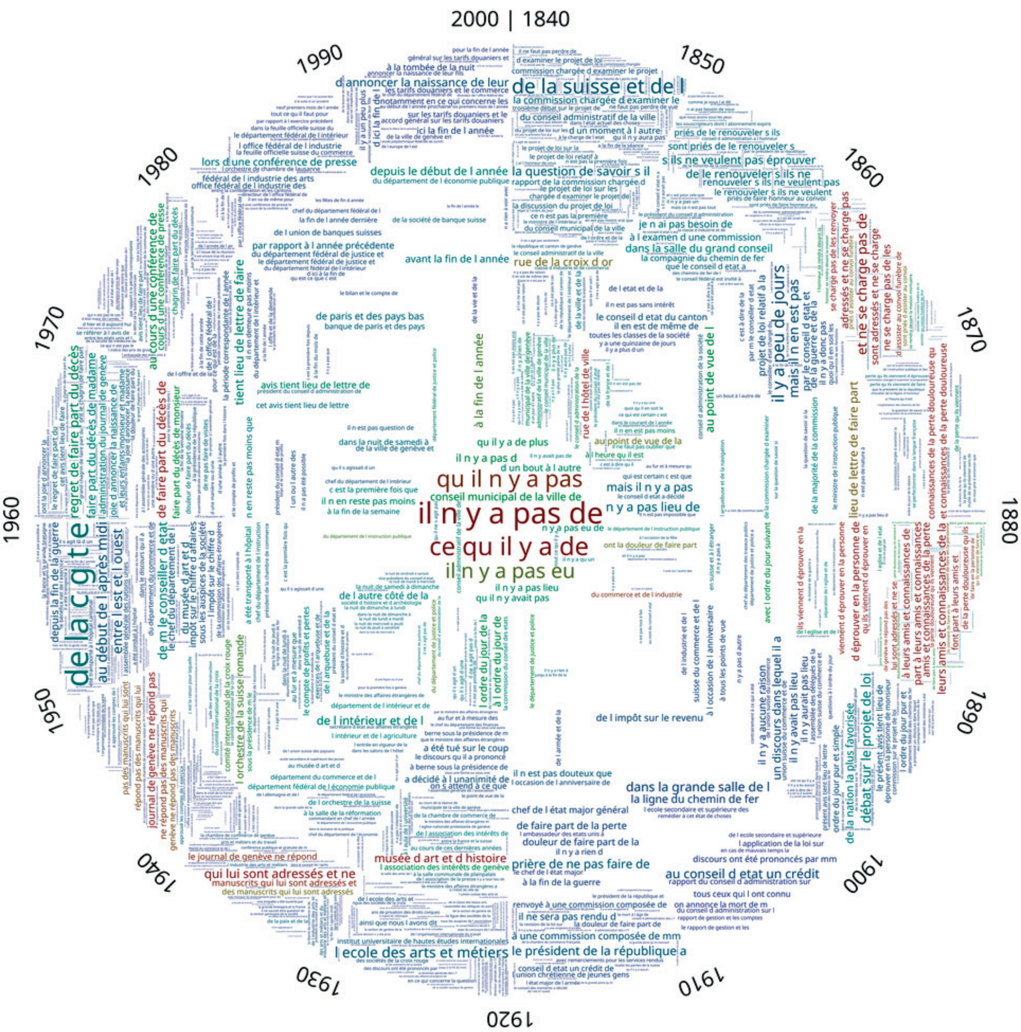


FIGURE 12.31 – Chronocloud classique sur les 6-grammes pour JDG

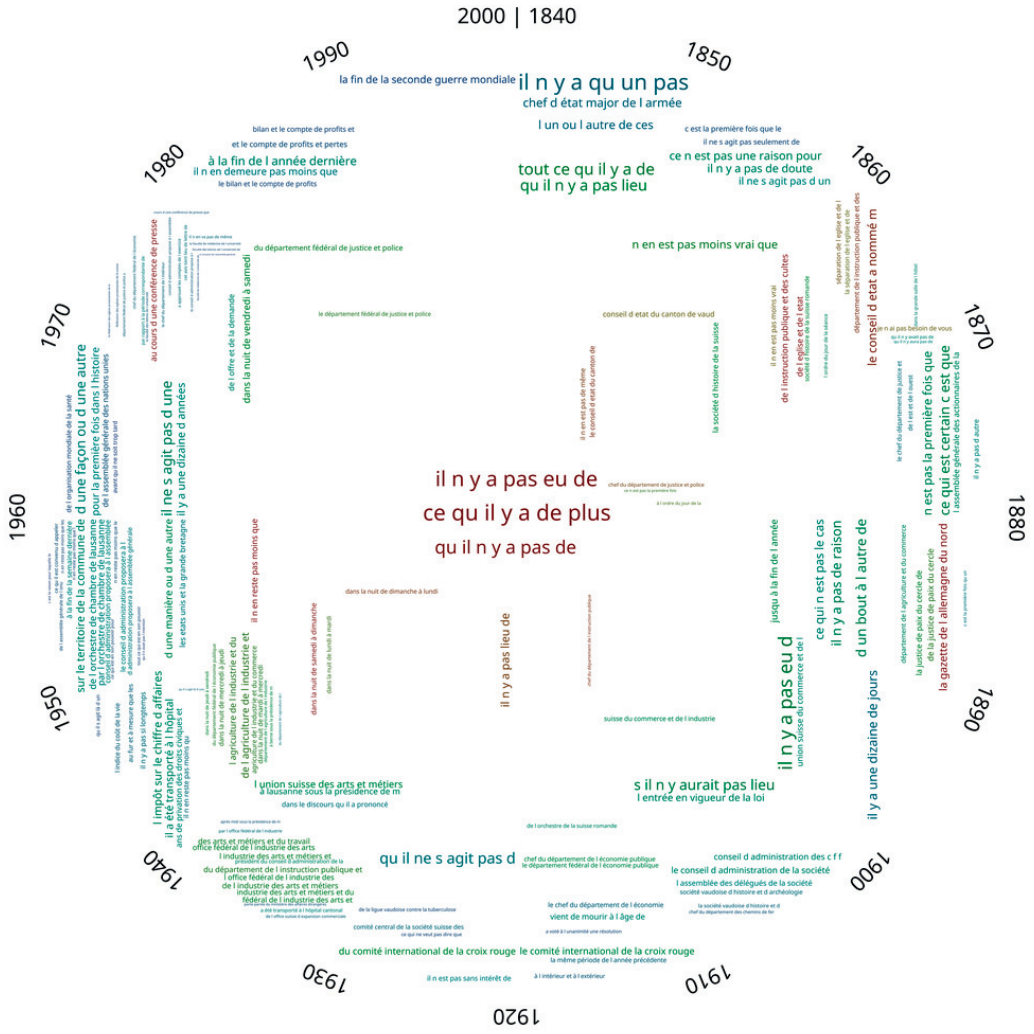


FIGURE 12.32 – Chronocloud classique sur les 7-grammes pour GDL

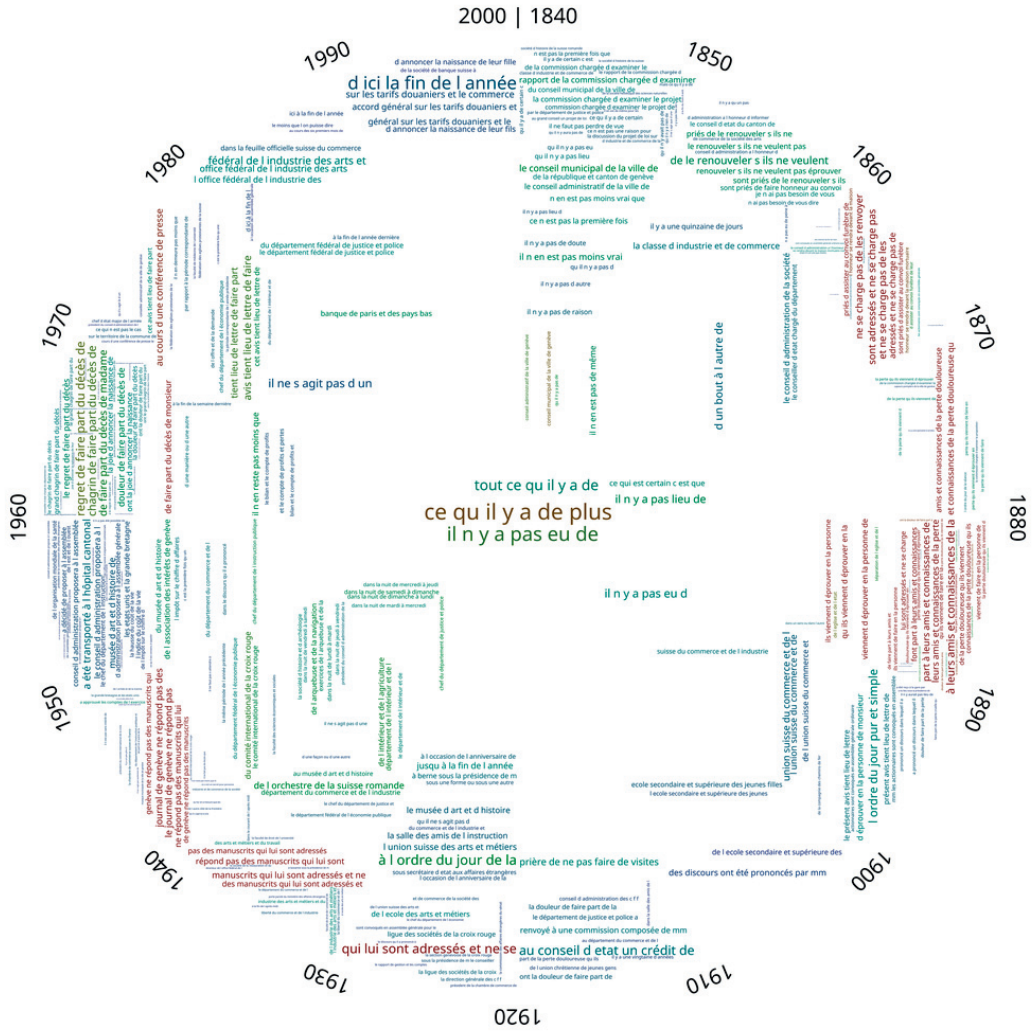


FIGURE 12.33 – Chronocloud classique sur les 7-grammes pour JDG

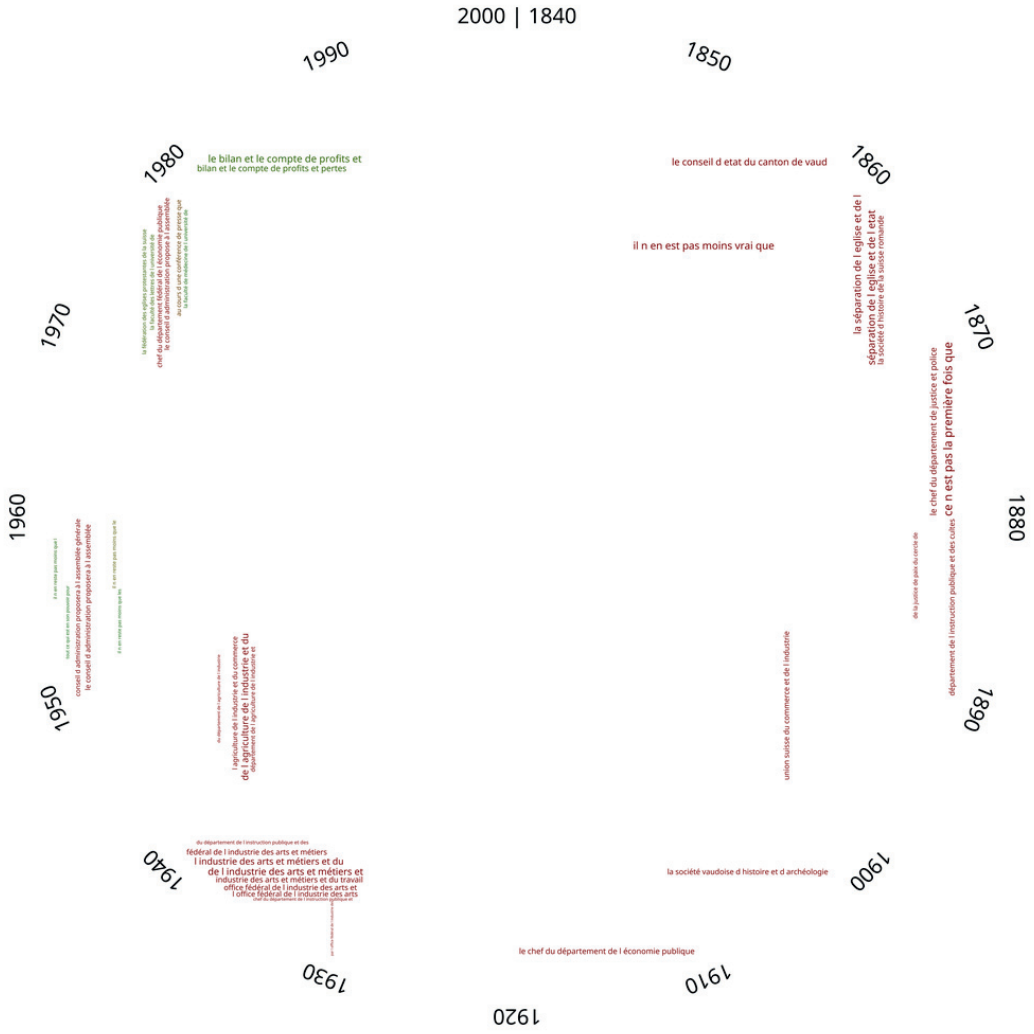


FIGURE 12.34 – Chronocloud classique sur les 8-grammes pour GDL

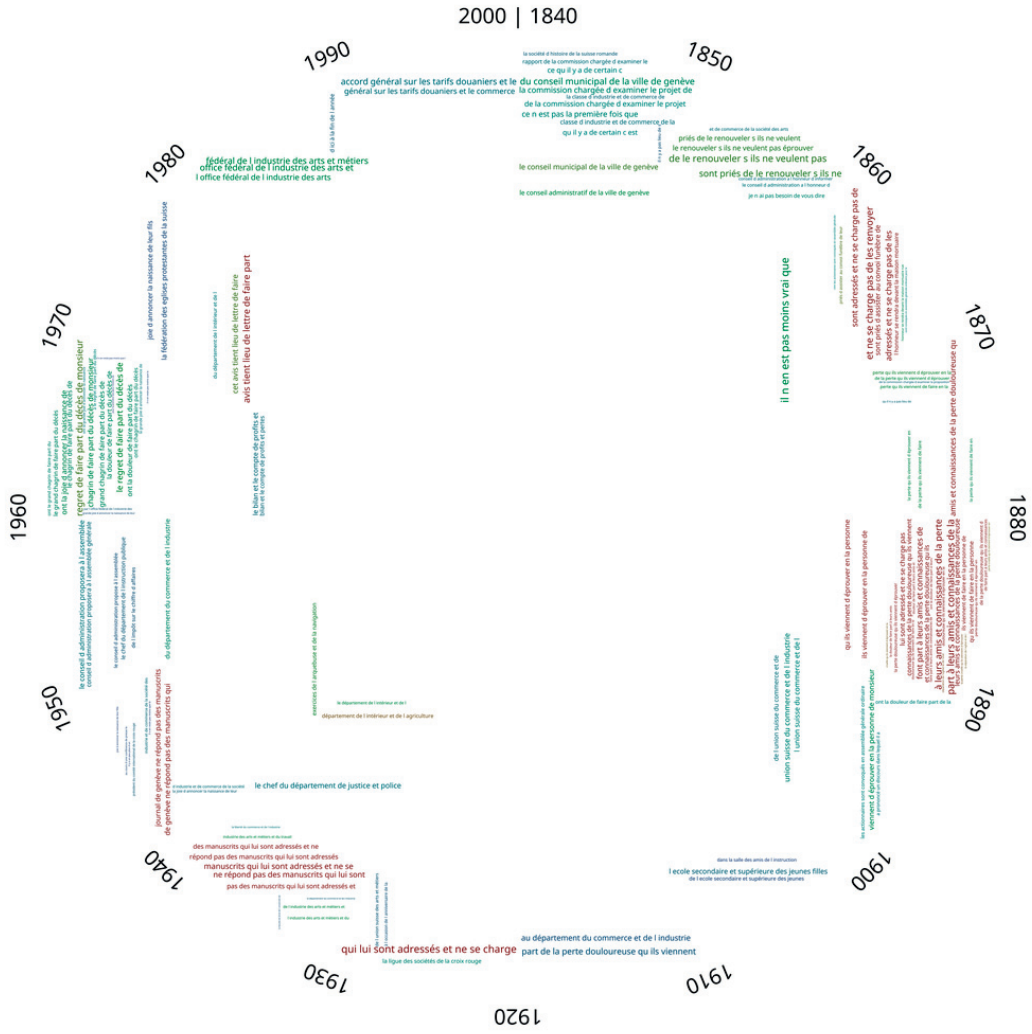


FIGURE 12.35 – Chronocloud classique sur les 8-grammes pour JDG

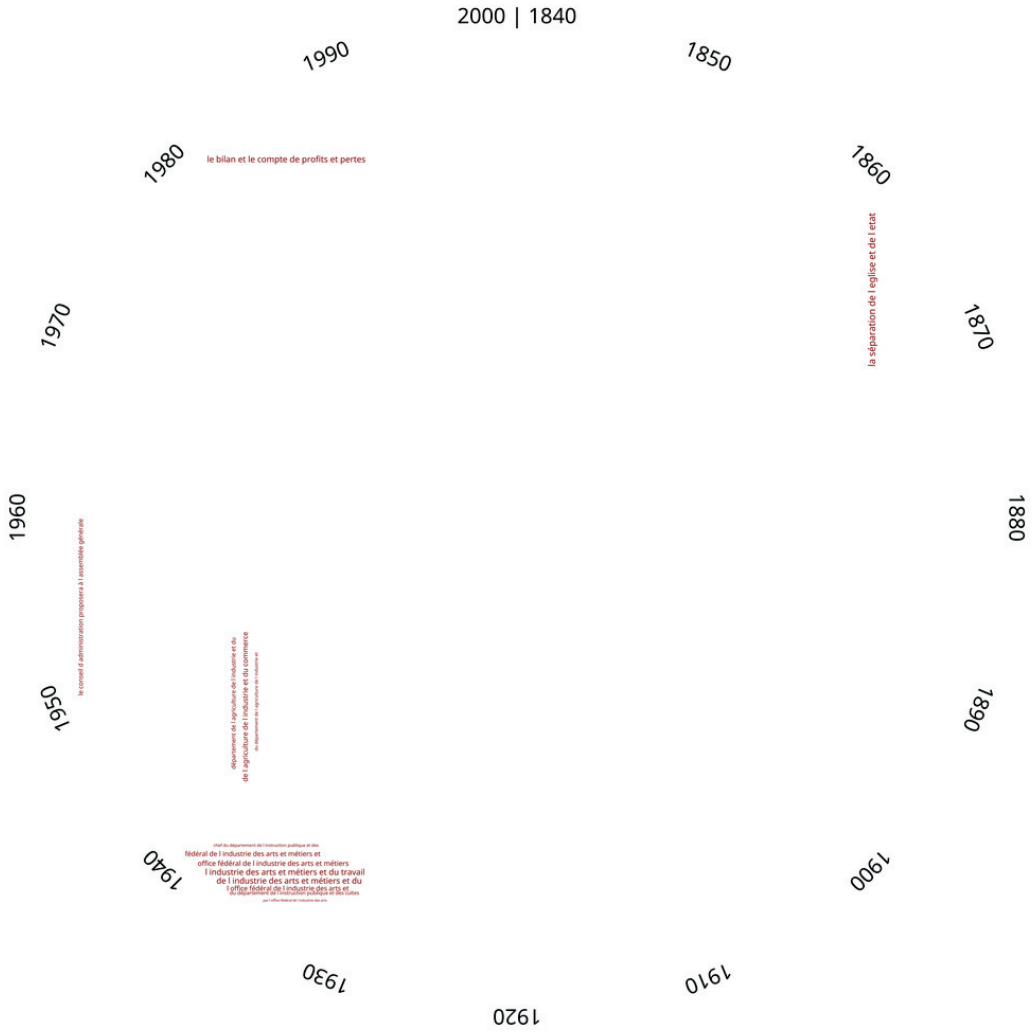


FIGURE 12.36 – Chronocloud classique sur les 9-grammes pour GDL

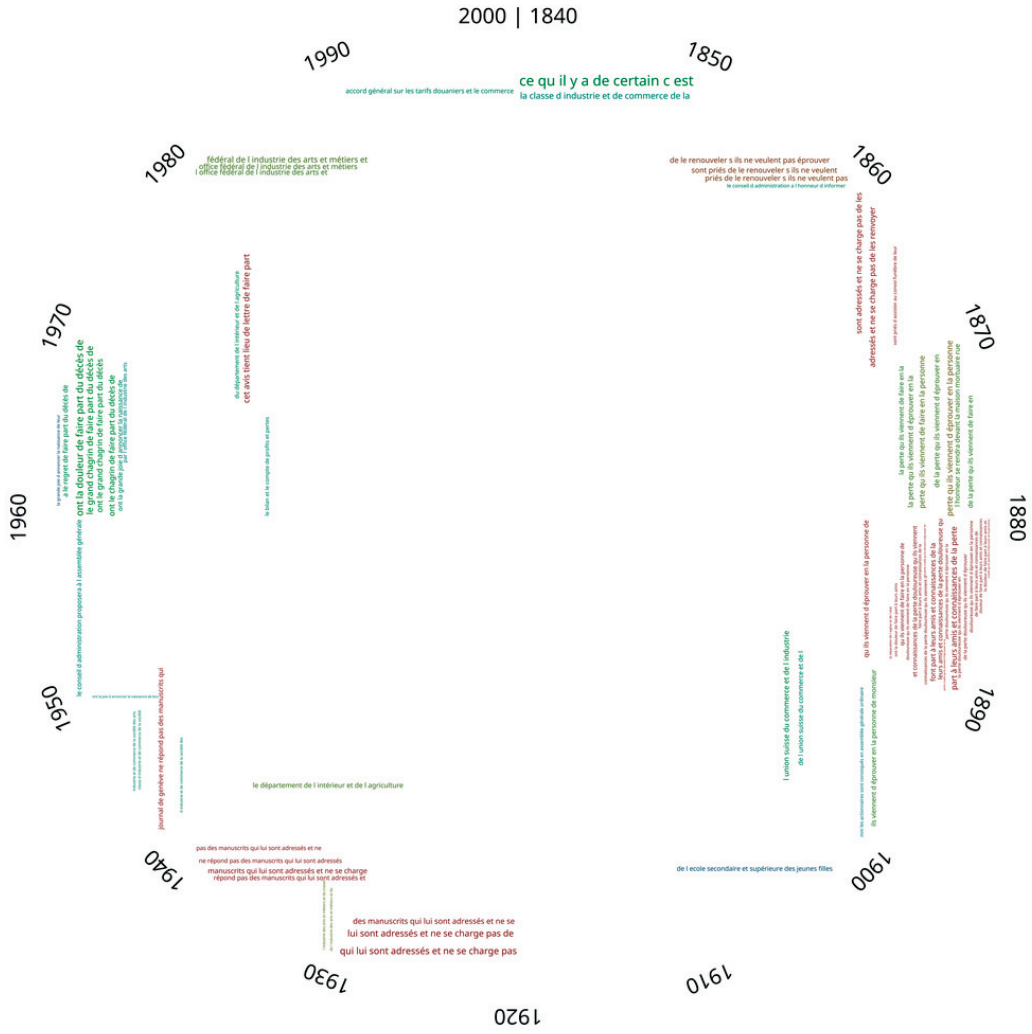


FIGURE 12.37 – Chronocloud classique sur les 9-grammes pour JDG

Ces chronoclouds sont particulièrement intéressants, car ils révèlent des combinaisons de mots autonomes tels que des combinaisons représentant des entités, bien souvent liées à la politique, au droit et aux institutions officielles. D'autres expressions sont plus liées à la langue elle-même comme certaines expressions "toutes faites" dont le sens global ne peut être directement déduit de celui des mots qui le composent. Ces combinaisons sont généralement appelées expression multi-mots (multiword expressions).

Quelques exemples de n-grammes intéressants révélés par les chronoclouds sont présentés dans les listes suivantes.

Entités politique ou juridique

La combinaison de mots forme une entité propre et l'ensemble se comporte de façon autonome. Par exemple, nous pouvons citer dans cette catégorie les n-grammes "office fédéral de l'industrie des arts et métiers", "chef du département de l'économie publique", "la faculté des lettres de l'université", "le musée d'art et d'histoire", "ligue des sociétés de la croix rouge", "assesseur de la justice de paix", "l'orchestre de la suisse romande", "chef de l'état major", "président de la république française", "président du conseil d'administration", "le ministère de l'intérieur", "le chancelier de l'empire", "l'armée de l'air", "commandement de l'armée", "conseil de l'Europe", "ministère de la guerre", "ministre de la défense", "le commandant en chef", "l'armée du salut", "grand duché de Bade", "l'empereur de Russie", "association des médecins", "la maison blanche", "le saint siège", "garde des sceaux", "affaire dreyfus", "parti communiste", "parti socialiste", "comptoir suisse" et "croix bleue".

Nous présentons dans la Figure 12.38, les profils fréquentiels des n-grammes "l'armée de l'air", "conseil de l'Europe", "ministère de la guerre" et "ministère de la défense".

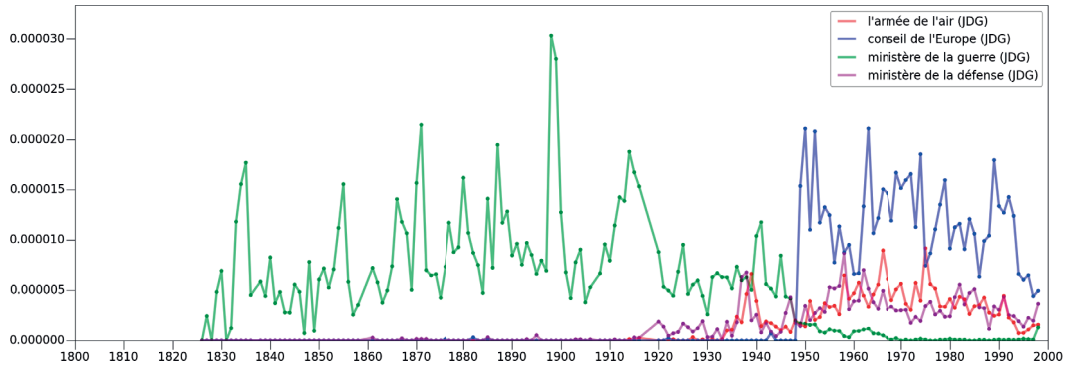


FIGURE 12.38 – Profil fréquentiel de "l'armée de l'air", "conseil de l'Europe", "ministère de la guerre" et "ministère de la défense" dans JDG

Le ministère de la guerre n'existe plus aujourd'hui et nous observons la décroissance de ce n-gramme jusqu'à la fréquence nulle alors qu'il fut le premier à être durablement présent

dans le corpus. Ensuite, le n-gramme "ministère de la défense" apparaît et prend le relais sous un nom moins offensif. Parallèlement à cela, le n-gramme "l'armée de l'air" est utilisé à une fréquence comparable. Enfin, le conseil de l'Europe fut créé par le traité de Londres après la guerre en 1949 et ce n-gramme apparaît donc brusquement dans le corpus dès 1949.

Concepts

Le n-gramme fait référence à un principe ou un concept. Par exemple, nous pouvons citer dans cette catégorie les n-grammes "la séparation de l'église et de l'état", "le bilan et le compte de profits et pertes", "l'indice du coût de la vie", "l'impôt sur le chiffre d'affaires", "liberté du commerce et de l'industrie", "l'offre et de la demande", "les droits de l'homme", "l'état de siège", "la liberté de conscience", "cessez le feu", "coupe du monde", "marque de fabrique", "indice des prix", "main d'oeuvre", "mise à prix", "mise au point", "mise sur pied", "déclaration de guerre", "les grandes lignes", "l'extrême droite", "l'extrême gauche", "hockey sur glace", "liberté et patrie", "tirage au sort", "groupe de travail", "prix Nobel", "porte-parole", "jeux olympiques", "contre attaque" et "main tendue".

Nous présentons dans la Figure 12.39, les profils fréquentiels des n-grammes "les droits de l'homme" ,"la liberté de conscience", "liberté et patrie" et "prix Nobel"

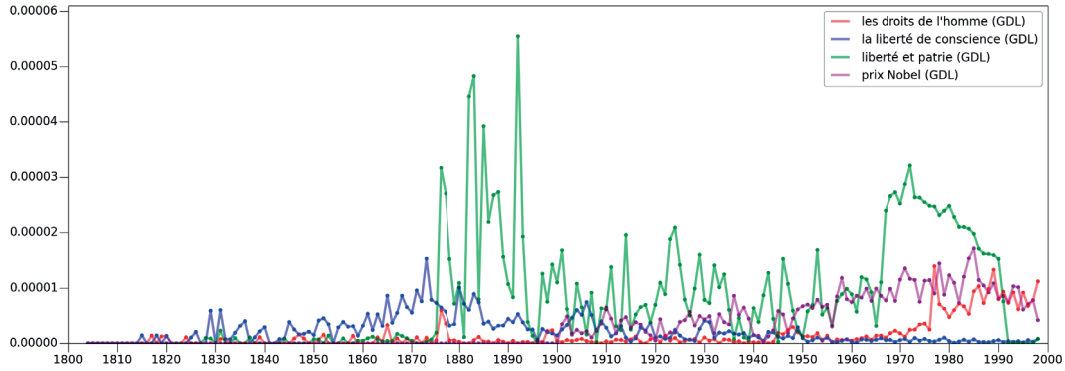


FIGURE 12.39 – Profil fréquentiel de "les droits de l'homme", "la liberté de conscience", "liberté et patrie" et "prix Nobel" dans GDL

Le n-.gramme "la liberté de conscience" a connu un pic fréquentiel en 1873 et est en forte diminution depuis. Toutefois, la devise de l'état de Vaud "liberté et patrie" a une grande variabilité fréquentielle, mais atteint la plus haute fréquence par rapport aux trois autres n-grammes. Elle n'est plus utilisée dès 1991, année de fusion de GDL avec JDG. Le n-gramme "prix Nobel" est apparu en 1901 peu après la création de la fondation Nobel, sa fréquence est ensuite en augmentation jusque 1985 puis en diminution. Enfin, le n-gramme "les droits de l'homme" est apparu plus tardivement, mais sa fréquence est en augmentation constante jusqu'à la fin du corpus.

Annonce officielle d'événements

Le n-gramme est repris tel quel, car il s'agit pour la plupart d'annonces formelles et/ou officielles disposant de formulations toute faites. Par exemple, nous pouvons citer dans cette catégorie les n-grammes "ont le grand chagrin de faire part du décès", "ont la grande joie d'annoncer la naissance", "le conseil d'administration a l'honneur d'informer", "cet avis tient lieu de lettre de faire part", "à leurs amis et connaissances de la perte douloureuse", "le regret de faire part du décès", "les actionnaires sont convoqués en assemblée générale ordinaire", "a décidé à l'unanimité de", "j'ai l'honneur de vous", "questions à l'ordre du jour", "ordre du jour de la séance", "le discours qu'il a prononcé", "joie d'annoncer la naissance", "la séance d'hier", "l'ordre du jour", "le présent avis", "tous droits réservés", "voici les résultats", "voix de majorité" et "avis mortuaires".

Nous présentons dans la Figure 12.40, les profils fréquentiels des n-grammes "ordre du jour de la séance", "le conseil d'administration a l'honneur d'informer", "ont le grand chagrin de faire part du décès" et "ont la grande joie d'annoncer la naissance".

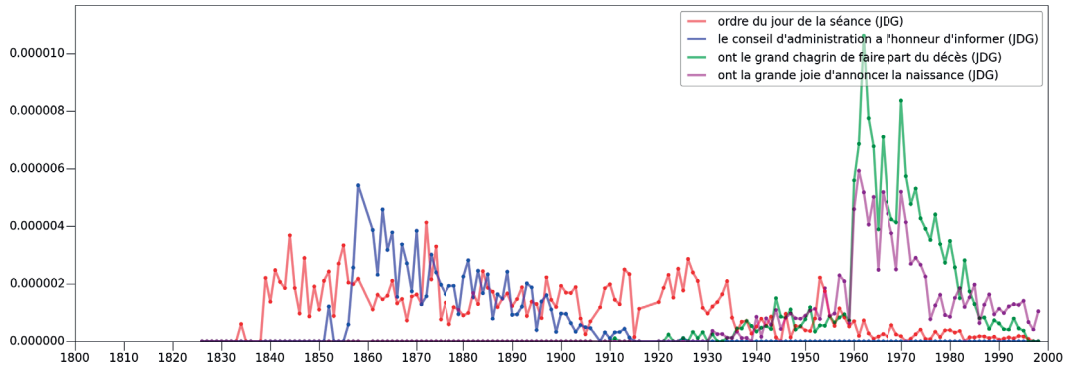


FIGURE 12.40 – Profil fréquentiel de "ordre du jour de la séance", "le conseil d'administration a l'honneur d'informer", "ont le grand chagrin de faire part du décès" et "ont la grande joie d'annoncer la naissance" dans JDG

Les deux premiers, "ordre du jour de la séance" et "le conseil d'administration a l'honneur d'informer", font partie d'un langage formel notamment utilisé par les entreprises et ils apparaissent dans les années les plus anciennes du corpus. Les deux derniers, "ont le grand chagrin de faire part du décès" et "ont la grande joie d'annoncer la naissance" apparaissent dans les années les plus récentes du corpus.

Expressions géographiques

Le n-gramme représente un lieu avec plus ou moins de précision. Par exemple, nous pouvons citer dans cette catégorie les n-grammes "en suisse et à l'étranger", "sur le territoire de la commune", "au bord de la mer", "dans le canton de Berne", "dans le canton de Fribourg",

"dans le canton de Vaud", "sur les bords du lac", "La Chaux-de-Fonds", "la Tour-de-Peilz", "la ville de Genève", "la ville de Lausanne", "l'Afrique du Sud", "à la policlinique", "à l'école", "à l'exposition", "Allemagne du nord", "pays de l'est", "au proche orient", "casino de Montbenon", "Château d'Oex", "dans la région", "dans les Balkans", "palais de Rumine", "pays de Vaud", "place St-François", "palais Bourbon", "Pays-Bas", "Etats-Unis", "Europe centrale" et "Tour Eiffel".

Nous présentons dans la Figure 12.41, les profils fréquentiels des n-grammes "palais de Rumine", "Europe centrale", "à l'école" et "à l'exposition".

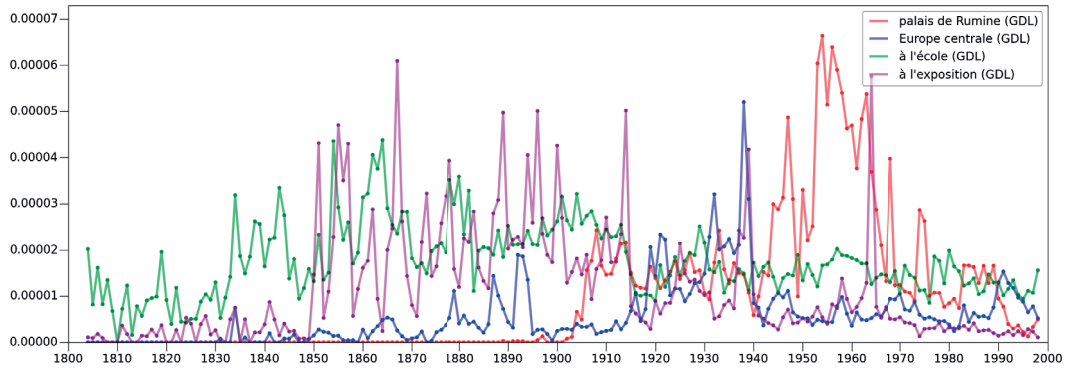


FIGURE 12.41 – Profil fréquentiel de "palais de Rumine", "Europe centrale", "à l'école" et "à l'exposition" dans GDL

Le palais de Rumine a été construit à la fin de XIX siècle ce qui explique l'apparition soudaine du n-gramme correspondant dans le corpus dès 1900. Dans la même zone de temps le terme "Europe centrale" apparaît. Les deux autres n-grammes sont moins précis et sont présents le long du corpus. Nous observons toutefois que le n-gramme "à l'exposition" possède des pics fréquentiels importants, dus au caractère fortement localisé dans le temps d'une exposition.

Expressions temporelles

Le n-gramme donne des informations temporelles avec plus ou moins de précision. Par exemple, nous pouvons citer dans cette catégorie les n-grammes "d'ici la fin de l'année", "à la fin de la semaine dernière", "avant qu'il ne soit trop tard", "il n'y a pas si longtemps", "dans la nuit de mardi à mercredi", "pour la première fois dans l'histoire", "il y a une dizaine de jours", "à la tombée de la nuit", "à l'heure qu'il est", "au cours de ces dernières années", "dans le courant du mois de", "d'un moment à l'autre", "tout au long de l'année", "il est encore trop tôt pour", "il y a peu de temps", "à l'époque de la", "au cours de la discussion", "au cours de la nuit", "il y a cinq ans", "il y a longtemps que", "il y a quelques jours", "dans la journée du", "à court terme", "à long terme", "contre la montre", "pour l'instant", "tout au long", "tout d'abord", "week-end" et "jours fériés".

Nous présentons dans la Figure 12.42, les profils fréquentiels des n-grammes "au cours de ces dernières années", "il y a quelques jours", "tout au long" et "pour l'instant".

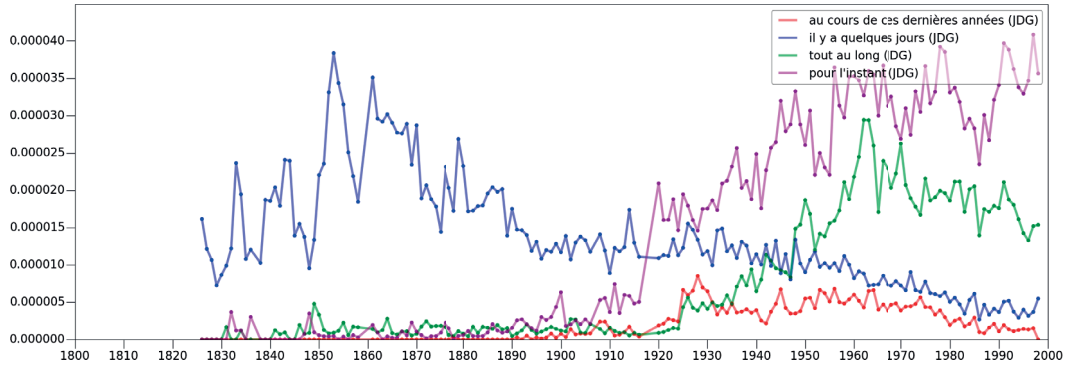


FIGURE 12.42 – Profil fréquentiel de "au cours de ces dernières années", "il y a quelques jours", "tout au long" et "pour l'instant" dans JDG

Nous observons sur les années les plus récentes que les expressions "au cours de ces dernières années" et "il y a quelques jours" sont en diminution tandis que les 3-grammes "tout au long" et "pour l'instant" sont en augmentation, hormis sur les années les plus récentes où "tout au long" diminue légèrement.

Toutefois, "tout au long" et "pour l'instant" sont apparus brusquement dans le corpus alors que le 6-gramme "au cours de ces dernières années" est apparu progressivement pour ensuite diminuer tout aussi progressivement que le 5-gramme "il y a quelques jours".

De façon surprenante, nous constatons que les expressions temporelles sont très nombreuses dans le chronocloud et elles possèdent un grand nombre de variantes selon le degré de précision voulu. Ces variantes ont des évolutions différentes, certaines plus anciennes sont en train de diminuer de fréquence tandis que d'autres formes apparaissent dans les années les plus récentes et augmentent en fréquence de façon importante.

La catégorie des expressions temporelles est particulièrement intéressante, car ces n-grammes sont adaptés à un grand nombre de situations et contextes différents. Les diverses variantes sont également présentes parmi les n-grammes les plus résilients et ceux-ci prennent une importance considérable dans les corpus de JDG et GDL.

Globalement, il est intéressant de constater que la dimension temporelle a une importance considérable dans le langage de la presse. L'abondance de ces expressions nous permet de relever la diversité permise par la langue française en matière de localisation dans le temps et cette diversité d'expressions temporelles semble toujours en mouvement.

L'étude de cette catégorie nous rapproche de notre objectif visant à isoler et à étudier les évolutions liées à la langue plutôt que diverses fluctuations dus à des événements particuliers.

Tournures de phrase

Le n-gramme représente une expression toute faite réutilisable dans de nombreuses et diverses situations. Par exemple, nous pouvons citer dans cette catégorie les n-grammes "ce qu'il y a de certain c'est", "ce n'est pas la première fois que", "il n'en est pas moins vrai que", "il n'en reste pas moins que", "d'une façon ou d'une autre", "d'une manière ou d'une autre", "n'en est pas moins vrai que", "il n'y a qu'un pas", "le moins que l'on puisse dire", "à tous les points de vue", "il n'est pas exclu que", "il n'est pas impossible que", "a rien à voir avec", "peu de chose près", "quoi qu'il en soit", "les uns et les autres", "c'est à dire", "d'ores et déjà", "de la part de", "de plus en plus", "de premier ordre", "par rapport à", "par rapport au", "point de vue", "relativement à la", "tout de même", "un peu partout", "sous le signe", "entrée en matière" et "fait l'objet".

Nous présentons dans la Figure 12.43, les profils fréquentiels des n-grammes "les uns et les autres", "d'ores et déjà", "sous le signe" et "tout de même".

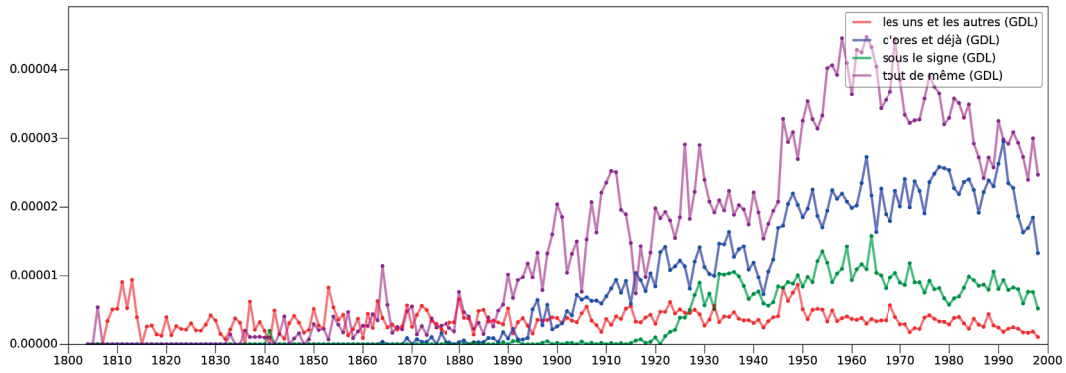


FIGURE 12.43 – Profil fréquentiel de "les uns et les autres", "d'ores et déjà", "sous le signe" et "tout de même" dans GDL

L'expression la plus ancienne "les uns et les autres" est stable et ne subit que peu de variations en comparaison avec les trois autres n-grammes. Toutefois, les n-grammes "tout de même", "d'ores et déjà" et "sous le signe" apparaissent respectivement en 1833, 1869, et 1917.

Ceux-ci se maintiennent durablement dans le corpus une fois que l'expression a atteint une fréquence stable. Le 3-gramme "tout de même" accuse une légère diminution dans les années les plus récentes.

Les éléments de cette catégorie sont particulièrement intéressants et font partie intégrante de l'analyse de la langue et de son fonctionnement. Ces tournures ont un sens spécifique et sont de nature à être réutilisées telles quelles dans diverses situations.

Elles ne respectent pas toujours une logique correcte de la langue et sont parfois renforcées simplement par l'usage de la masse qui l'adopte quand bien même la logique serait imprécise.

Synthèse

Le chronocloud a révélé des n-grammes que nous avons classifiés en catégories. Parmi ces catégories, nous relevons que les "expressions temporelles" et les "tournures de phrase" sont d'autant plus intéressantes qu'elles sont applicables à de nombreux contextes. En effet, nous pouvons considérer que l'apparition dans le corpus d'un n-gramme représentant une entité politique ou juridique est un effet qui n'influence pas durablement l'utilisation du langage dans d'autres contextes. En revanche, un n-gramme représentant une façon particulière d'arranger les mots pour être utilisable dans de nombreuses circonstances aura un impact plus important sur la langue et peut être considéré comme faisant partie de son fonctionnement de façon plus centrale.

En outre, nous observons que ces catégories sont sujettes à d'importantes variations fréquentielles et ce particulièrement sur les années les plus récentes. Durant cette période, de nombreux n-grammes apparaissent et leurs utilisations a pour effet d'augmenter leurs fréquences au détriment d'autres expressions plus anciennes. De façon intéressante, nous observons aussi que la plupart des noms communs et n-grammes formant une entité sont mis en évidence dans le chronocloud avec un déterminant permettant d'identifier la différence de genre de ces entités. Sur un échantillon de 1061 différents n-grammes repérés sur le chronocloud, nous présentons la répartition en terme de catégorie dans la Table 12.3.

	9	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**totaux**
Entités	4	13	16	26	44	56	113	62	334
Concepts	2	1	4	9	20	21	85	60	202
Annonces officielles	5	2	0	7	3	2	4	3	26
Expressions géographiques	0	0	0	5	7	9	23	17	61
Expressions temporelles	0	1	12	38	34	7	21	6	119
Tournures de phrase	1	3	26	82	76	30	52	49	319
Entités	1%	4%	5%	8%	13%	17%	34%	19%	31%
Concepts	1%	0%	2%	4%	10%	10%	42%	30%	19%
Annonces officielles	19%	8%	0%	27%	12%	8%	15%	12%	2%
Expressions géographiques	0%	0%	0%	8%	11%	15%	38%	28%	6%
Expressions temporelles	0%	1%	10%	32%	29%	6%	18%	5%	11%
Tournures de phrase	0%	1%	8%	26%	24%	9%	16%	15%	30%

TABLE 12.3 – Répartition absolue et relative des n-grammes par catégorie

Nous remarquons que les catégorie "Entités", "Concepts" et "Expressions géographiques" ont un taux plus important de 3-grammes alors que les catégories "Expressions temporelles" et "Tournures de phrase" incluent un taux plus important de 5-grammes et 6-grammes.

Chronocloud différentiel

Dans cette section nous présentons les chronoclouds différentiels asymétriques pour chaque niveau $n > 1$ et nous détaillons ensuite nos observations en considérant l'ensemble des chronoclouds différentiels asymétriques des différents niveaux. Vue la faible quantité de n-grammes communs aux niveaux supérieurs, nous considérons uniquement les niveaux n jusque $n = 5$, nous présentons 10 chronoclouds différentiels asymétriques dans les Figures 12.44 à 12.51.

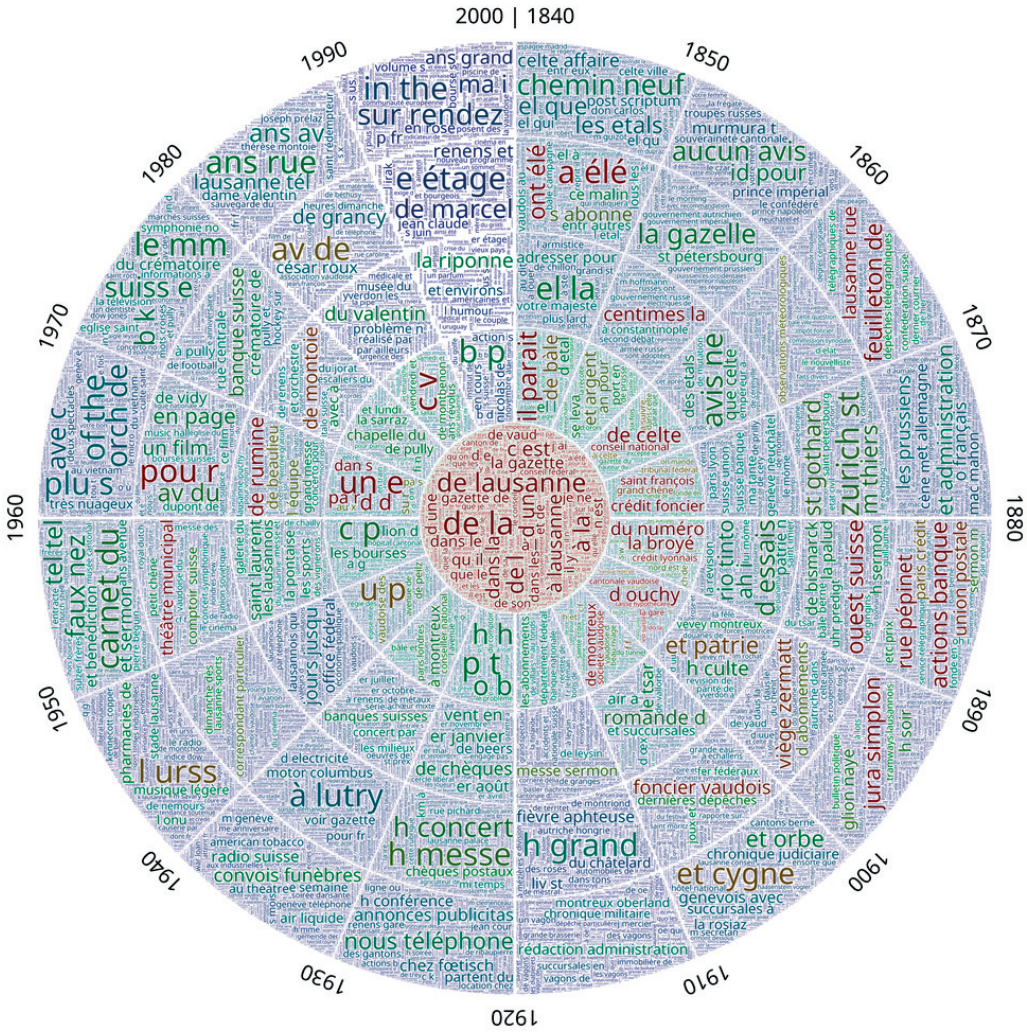


FIGURE 12.44 – Chronocloud différentiel asymétrique de 2-grammes du corpus de GDL moins celui de JDG

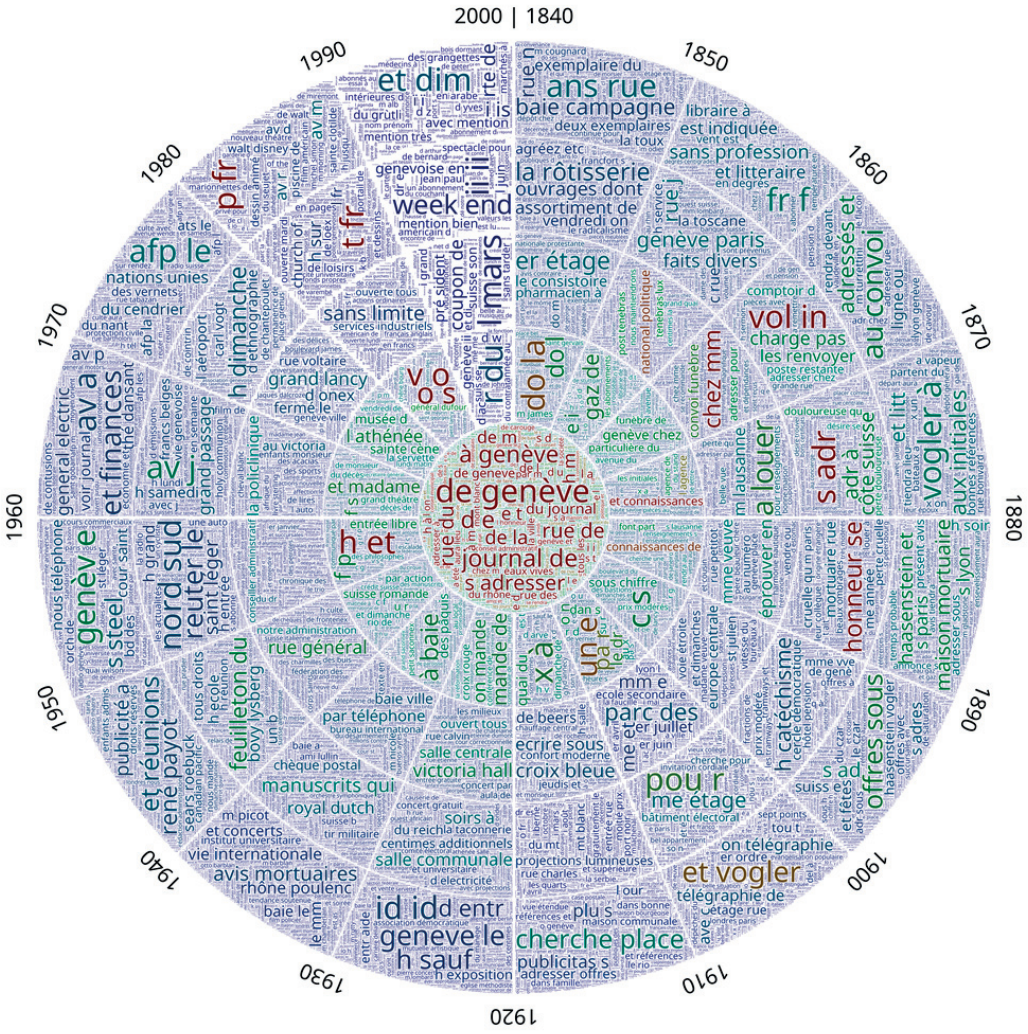


FIGURE 12.45 – Chronocloud différentiel asymétrique de 2-grammes du corpus de JDG moins celui de GDL

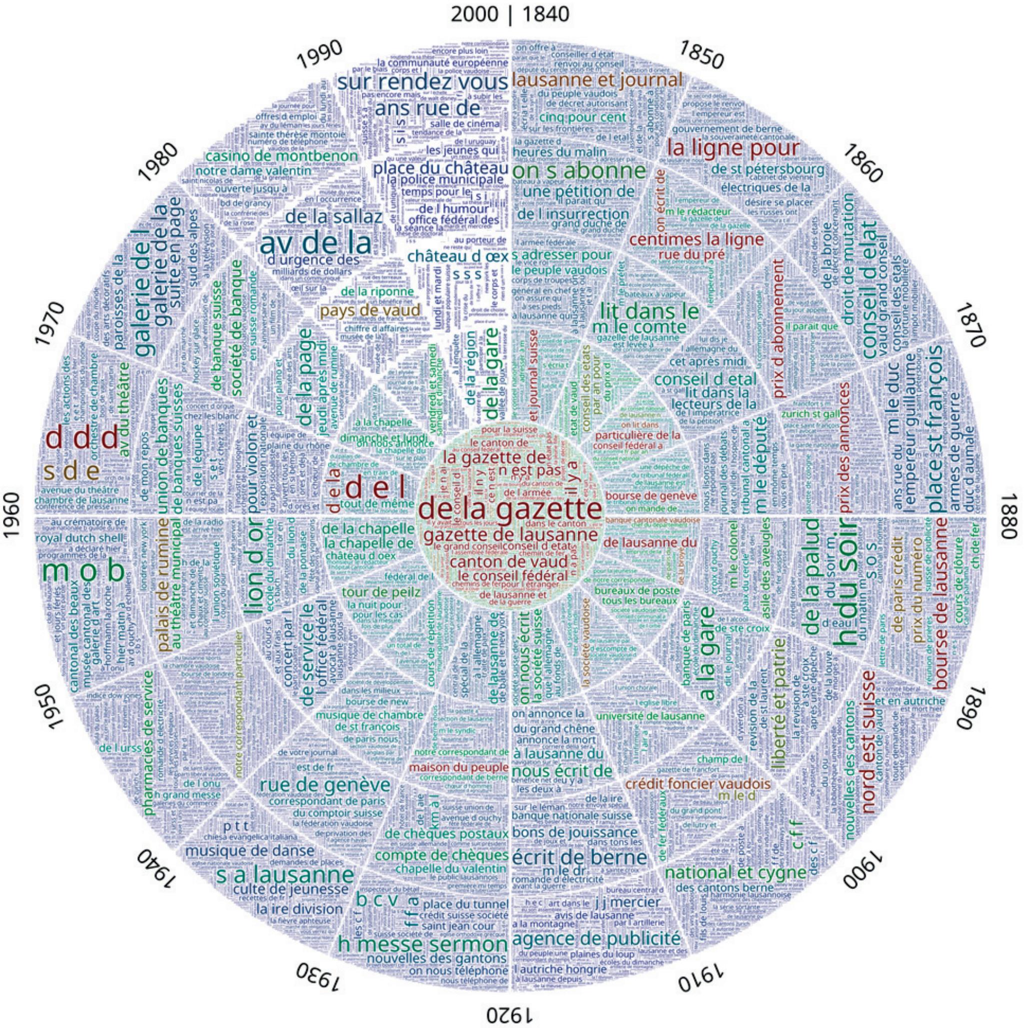


FIGURE 12.46 – Chronocloud différentiel asymétrique de 3-grammes du corpus de GDL moins celui de JDG

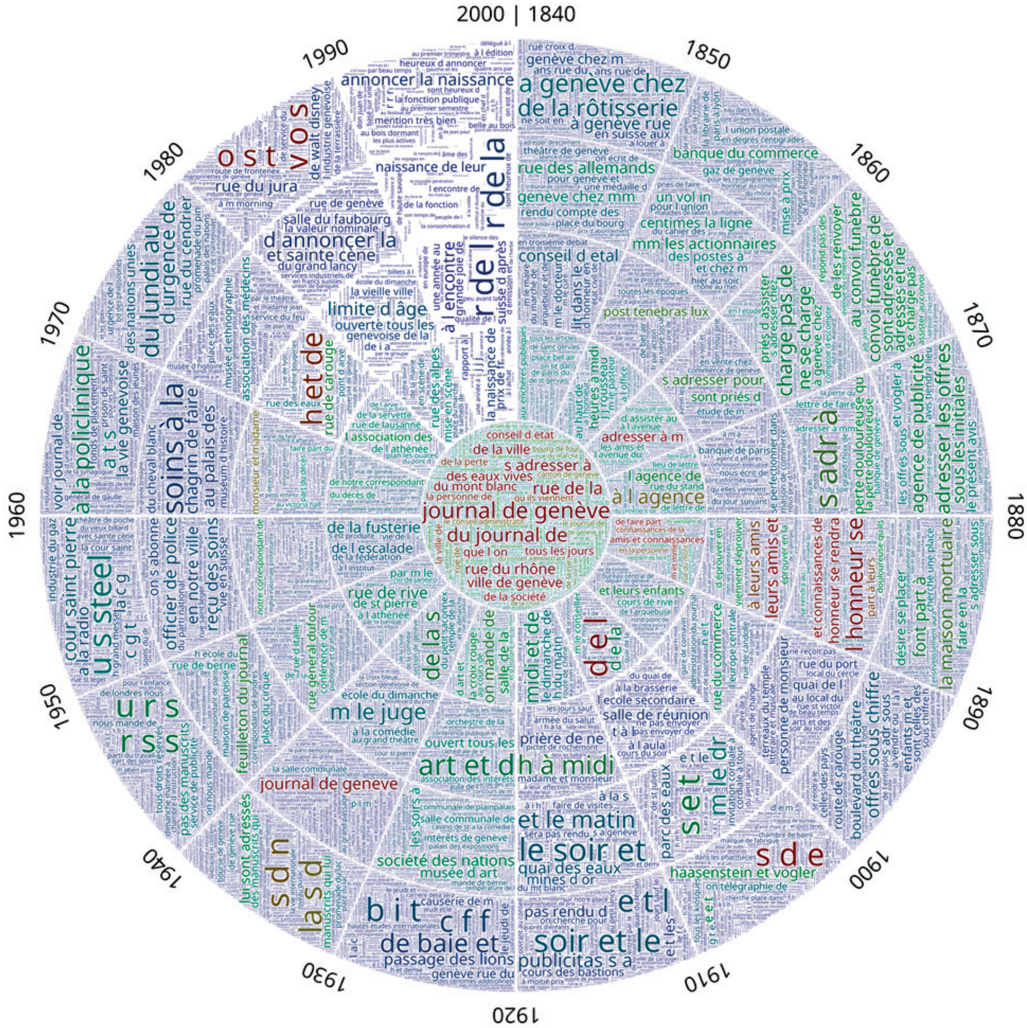


FIGURE 12.47 – Chronocloud différentiel asymétrique de 3-grammes du corpus de JDG moins celui de GDL

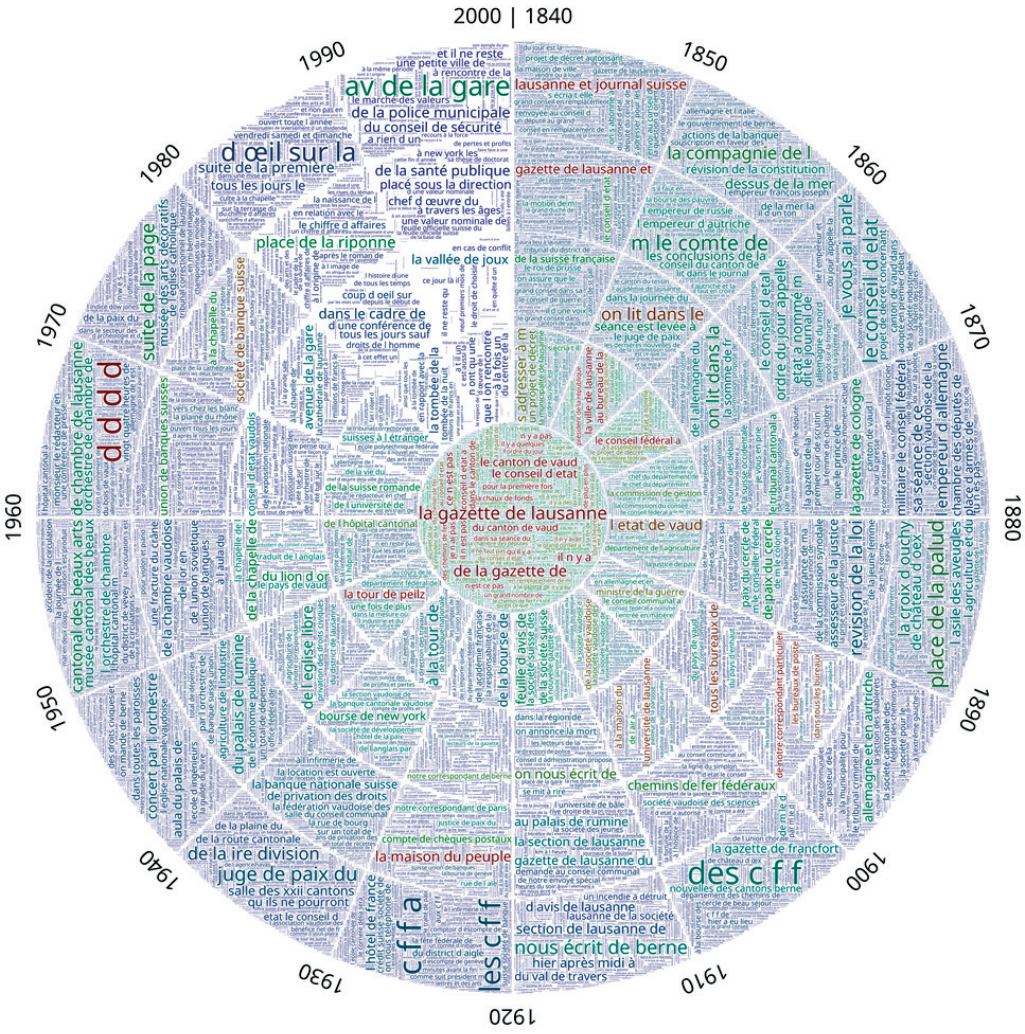


FIGURE 12.48 – Chronocloud différentiel asymétrique de 4-grammes du corpus de GDL moins celui de JDG

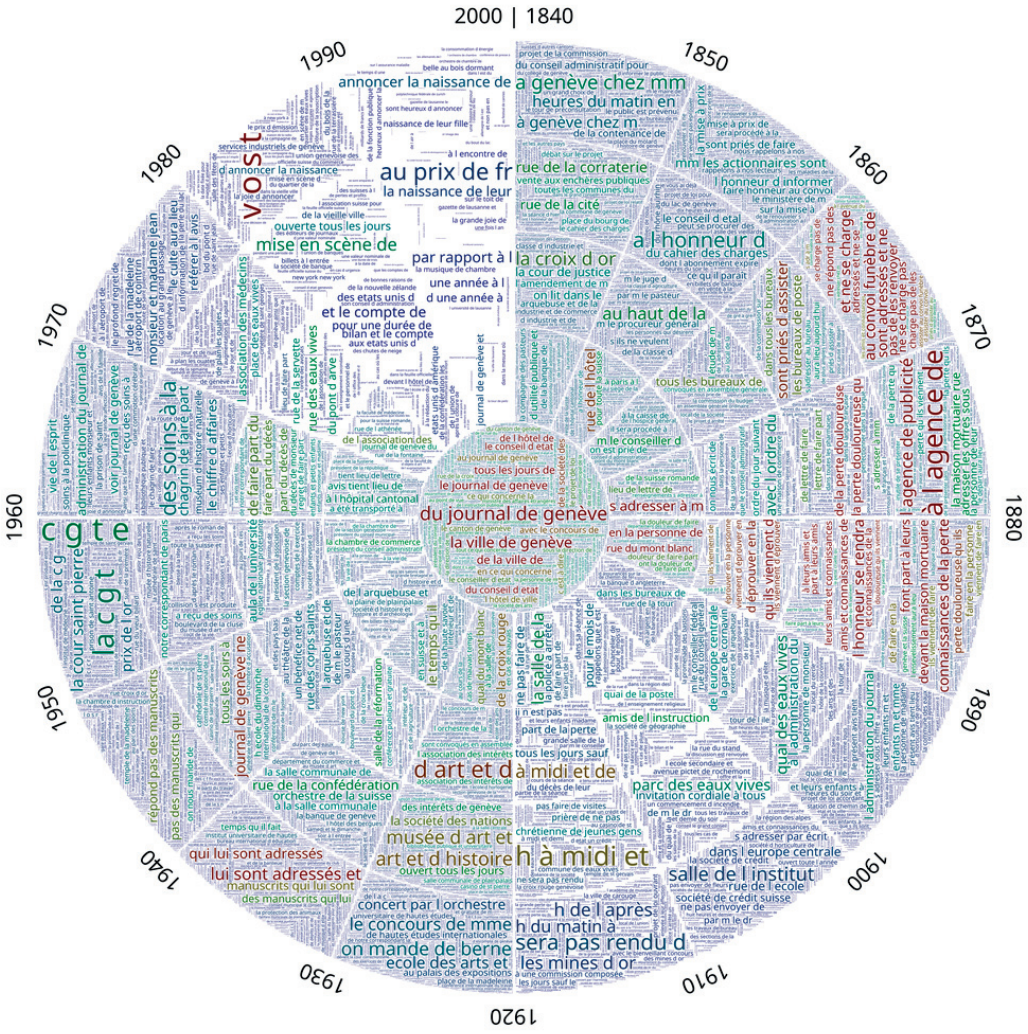


FIGURE 12.49 – Chronocloud différentiel asymétrique de 4-grammes du corpus de JDG moins celui de GDL

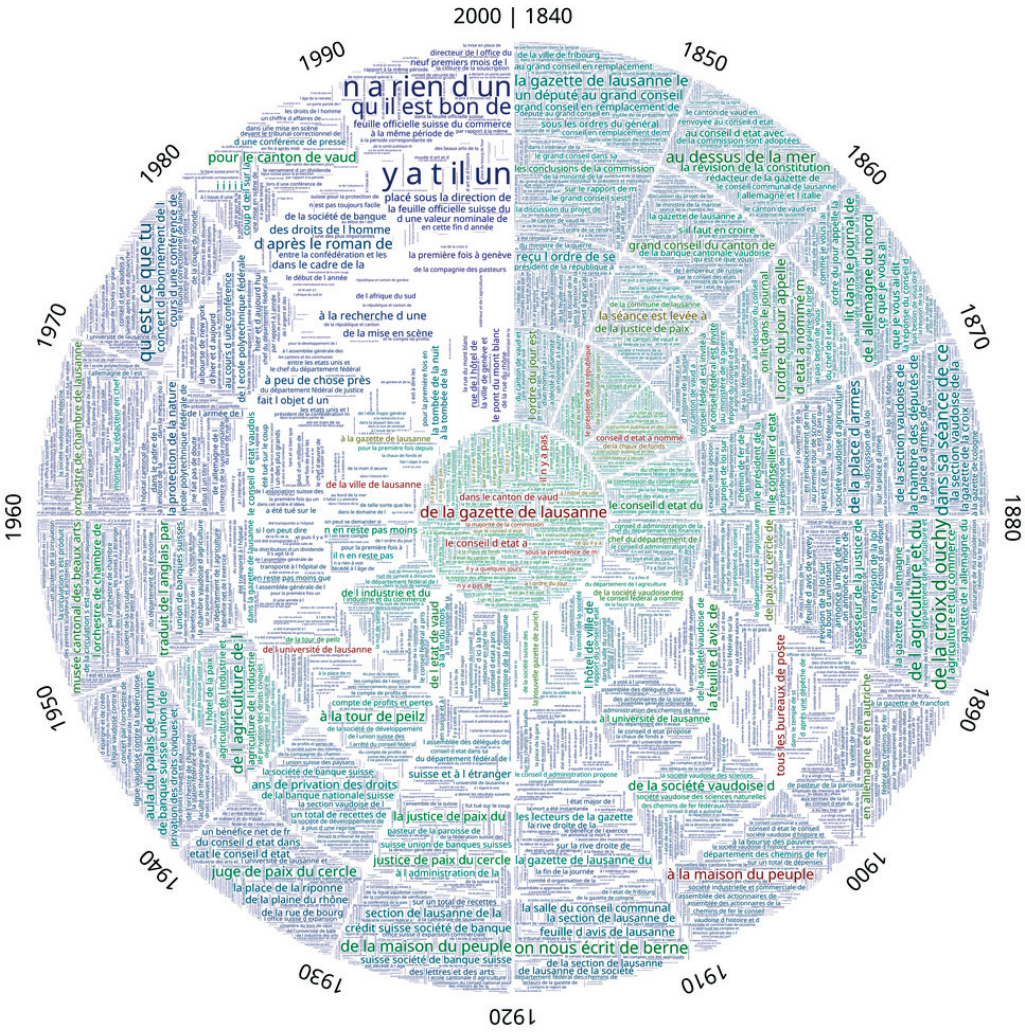


FIGURE 12.50 – Chronocloud différentiel asymétrique de 5-grammes du corpus de GDL moins celui de JDG

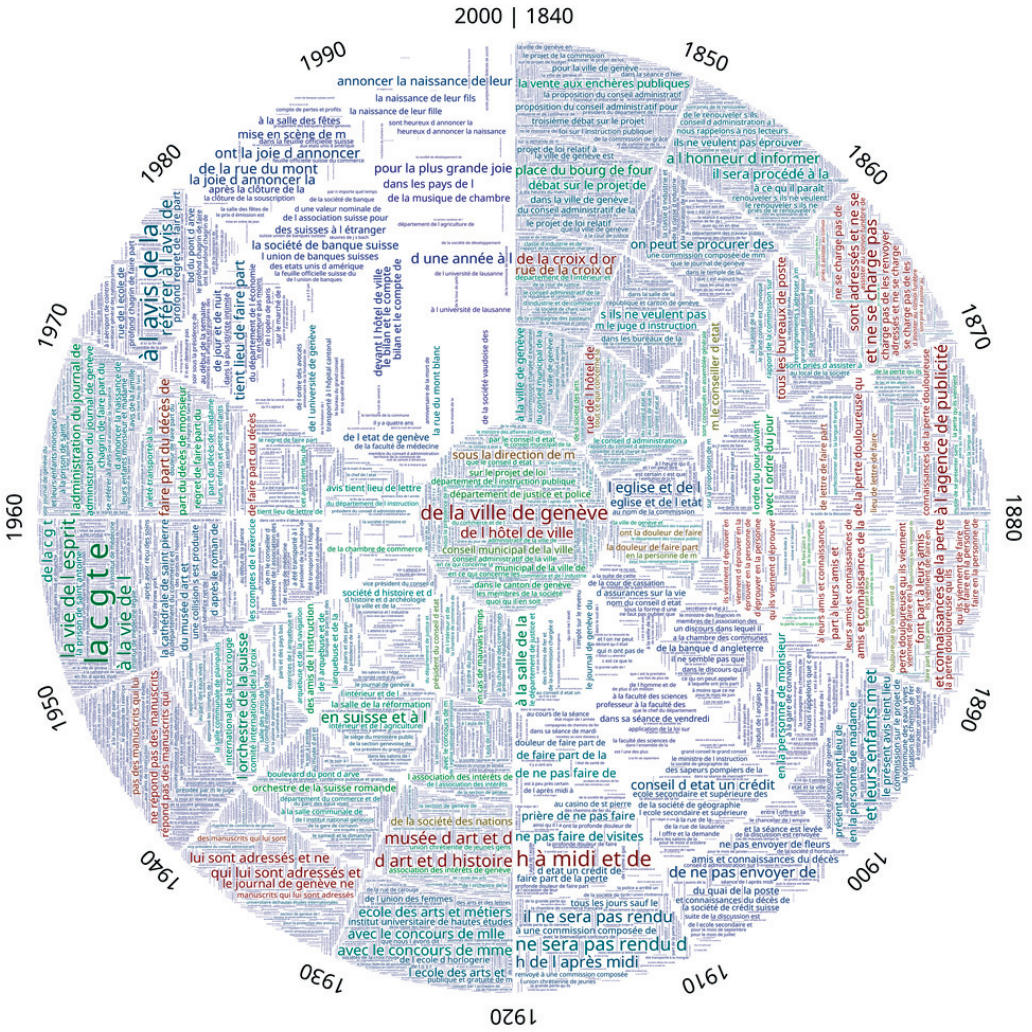


FIGURE 12.51 – Chronocloud différentiel asymétrique de 5-grammes du corpus de JDG moins celui de GDL

Comme dans le cas du niveau $n = 1$, un grand nombre de n-grammes visibles dans ces représentations sont des lieux géographiques, des adresses, des corporations ou des personnes connues, mais à un niveau local. Pour les mêmes raisons que le niveau des mots, la différence de localisation des journaux induit des différences d'évolution fréquentielle principalement au niveau des lieux et personnes rattachés à la région.

L'exemple canonique du 4-gramme "la ville de" suivi du nom "Lausanne" ou "Genève" est présenté dans la Figure 12.52.

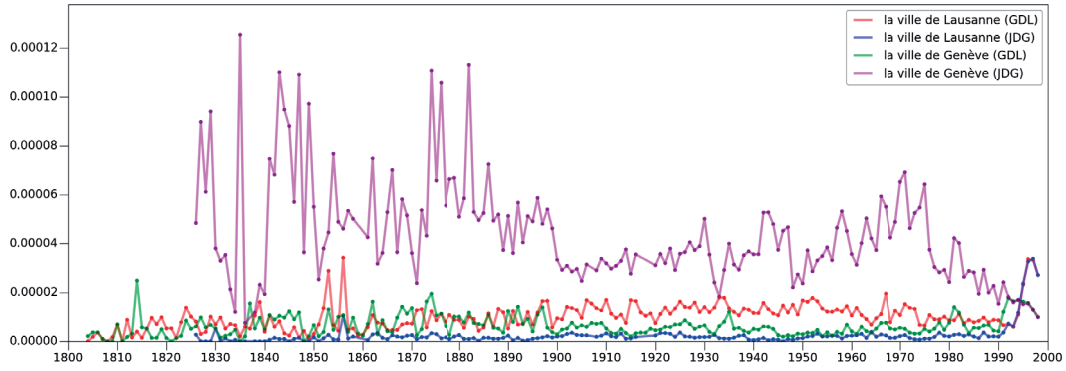


FIGURE 12.52 – Profils fréquentiels de "la ville de Lausanne" et "la ville de Genève"

Nous observons que "la ville de Lausanne" est très peu utilisé par JDG et inversement "la ville de Genève" est très peu utilisé par GDL. De façon intéressante, la fusion des deux journaux en 1991 semble utiliser le 4-gramme "la ville de Lausanne" bien plus que ne le faisaient les journaux séparés.

Afin d'élargir nos observations au niveau régional, nous présentons l'exemple des 4-grammes "La Tour-de-Peilz" et "La Chaux-de-Fonds" dans la Figure 12.53.

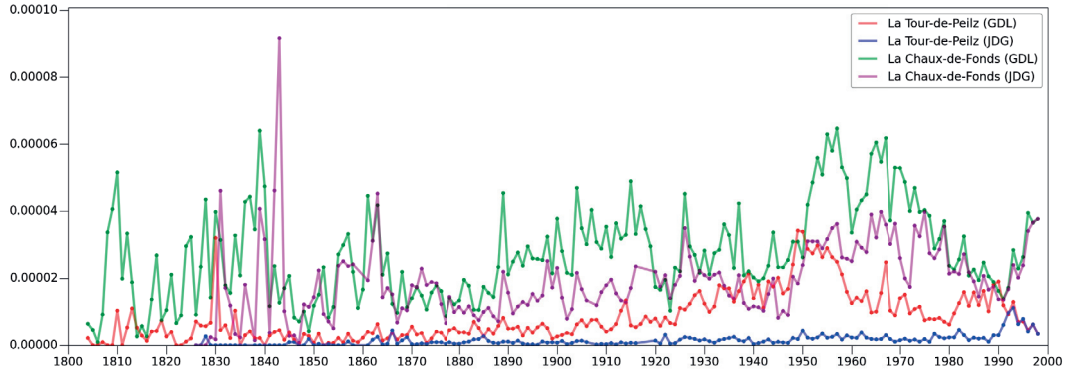


FIGURE 12.53 – Profils fréquentiels de "La Tour-de-Peilz" et "La Chaux-de-Fonds"

Nous observons des effets similaires sur les localisations de type régional. Les exemples de différences géographiques sont multiples dans ces corpus, sans compter les nombreuses adresses mises en évidence par les chronoclouds différentiels sous la forme de n-grammes commençant par le mot "rue". Toutefois, ces exemples ne font pas partie des exemples parmi les plus intéressants pour notre objectif vu que leur utilisation est totalement contextuelle.

Les chronoclouds révèlent plusieurs expressions rattachées par exemple au 5-gramme "à leurs amis et connaissances". Ce 5-gramme est exhibé dans les chronoclouds différentiels ainsi que dans les chronoclouds classiques. Dans les chronoclouds classiques, les n-grammes contenant "à leurs amis et connaissances" vont jusque au niveau $n = 9$ comme par exemple le 9-gramme "part à leurs amis et connaissances de la perte".

Il est clair qu'il s'agit d'une phrase entière qui est reprise de journal en journal afin d'annoncer le décès de personnes. Nous présentons les n-grammes "à leurs amis et connaissances" et "part à leurs amis et connaissances de la perte" dans la Figure 12.54.

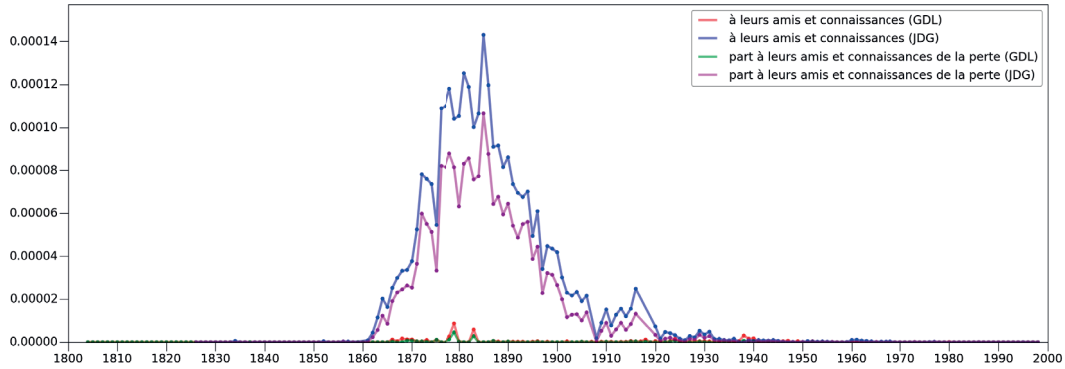


FIGURE 12.54 – Profils fréquentiels de "à leurs amis et connaissances" et "part à leurs amis et connaissances de la perte"

Nous observons que ces expressions ne sont jamais utilisées par GDL, mais elles sont utilisées durablement sur plus d'une cinquantaine d'année par JDG. Toutefois, cet exemple n'est pas issu d'une utilisation statistique de la langue, mais plutôt d'un choix éditorial de long terme.

Cependant, cet exemple apporte un éclairage sur la façon d'annoncer un événement parmi les plus sensibles, le décès d'une personne. Nous observons que la fréquence monte progressivement et puis descend à peu près de la même manière. L'usage de ces expressions évolue donc aussi et elles sont remplacées par d'autres formules consacrées.

Outre ces différences de corpus causées par une sur-représentation de certains n-grammes dans diverses sections particulières du journal et outre les n-grammes considérés comme externes au langage par définition comme les lieux, les noms de personne, les compagnies et corporations, le chronocloud différentiel montre quelques exemples d'usages linguistiques différents entre les deux journaux.

Par exemple, le n-grammes "Suisse romande" désigne la partie francophone de la Suisse. En réalité, l'expression la plus ancienne retrouvée dans le corpus est "Suisse française". Ces 2-grammes étant mis en évidence par le chronocloud différentiel, nous présentons leurs profils fréquentiels dans les Figures 12.55 et 12.56.

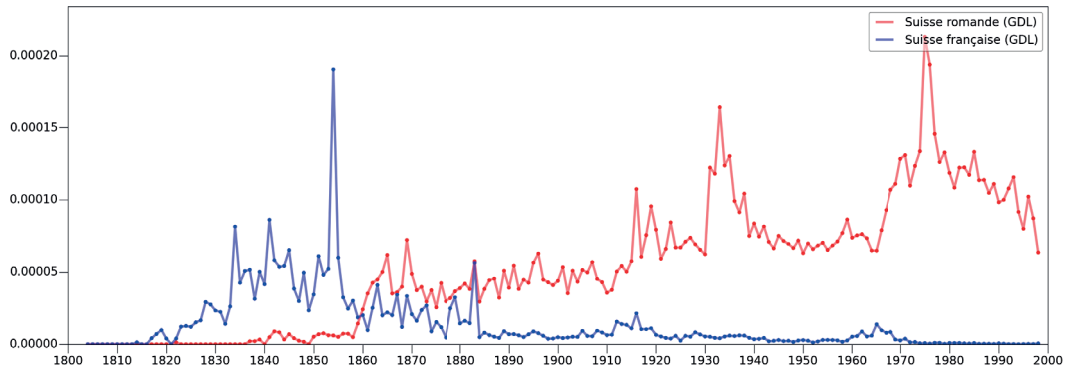


FIGURE 12.55 – Profils fréquentiels de "Suisse romande" et "Suisse française" pour GDL

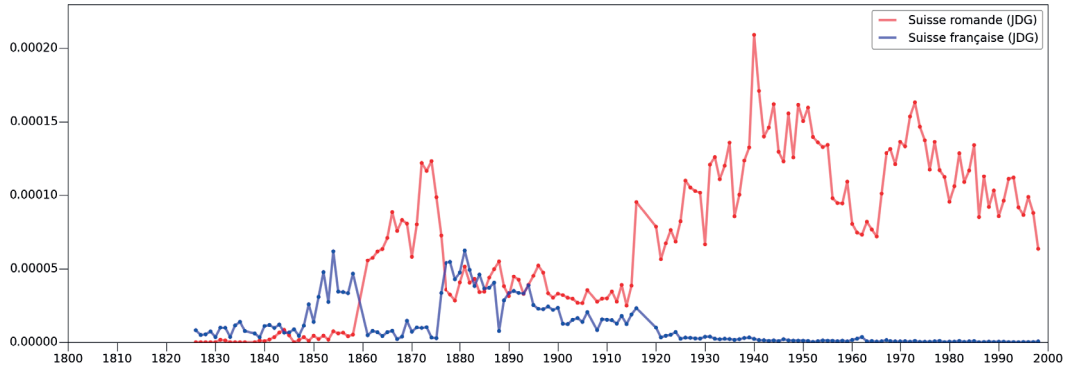


FIGURE 12.56 – Profils fréquentiels de "Suisse romande" et "Suisse française" pour JDG

Nous observons dans les deux cas que l'expression "Suisse française" a été utilisée en premier. Le 2-gramme "Suisse romande" apparaît ensuite en 1860. L'effet de cette apparition est une diminution progressive de la fréquence de "Suisse française" pour GDL, mais par contre JDG n'utilise brusquement plus cette expression.

Etonnamment, le 2-gramme "Suisse française" revient dans le cas de JDG en 1875 et a ensuite une fréquence similaire à "Suisse romande" jusque 1920 alors que GDL n'utilise quasiment plus cette expression déjà depuis 1885.

Le résultat à long terme est le même puisque seule l'expression "Suisse romande" est utilisée aujourd'hui, mais cet exemple nous montre un effet de "lutte" entre les deux 2-grammes qui ont une histoire différente dans chaque journal.

Un autre exemple de ce type peut être repéré sur le chronocloud différentiel avec le 2-gramme "dépêches télégraphiques" utilisé par les journaux pour qualifier une dépêche par télégramme. Le profil fréquentiel de "dépêches télégraphiques" est à mettre en parallèle avec celui de "dépêches électriques" qui était également utilisé par GDL et JDG dans la seconde moitié du XIX siècle. Nous présentons donc leurs profils fréquentiels dans les Figures 12.57 et 12.58.

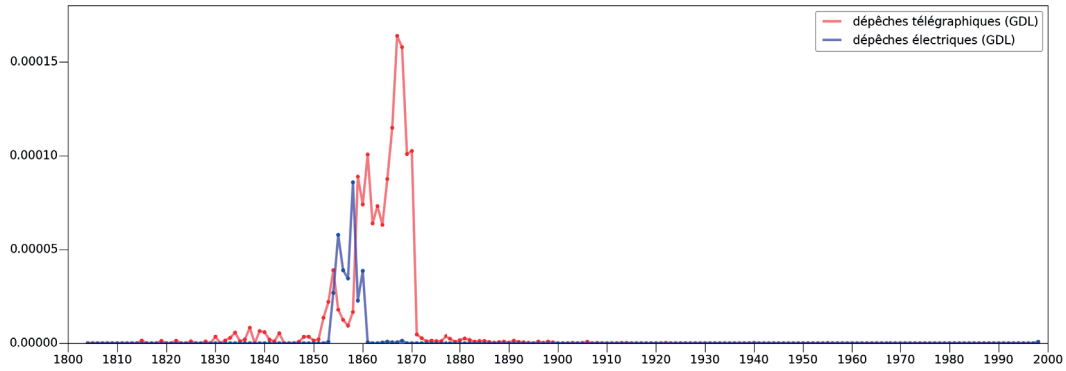


FIGURE 12.57 – Profils fréquentiels de "dépêches télégraphiques" et "dépêches électriques" pour GDL

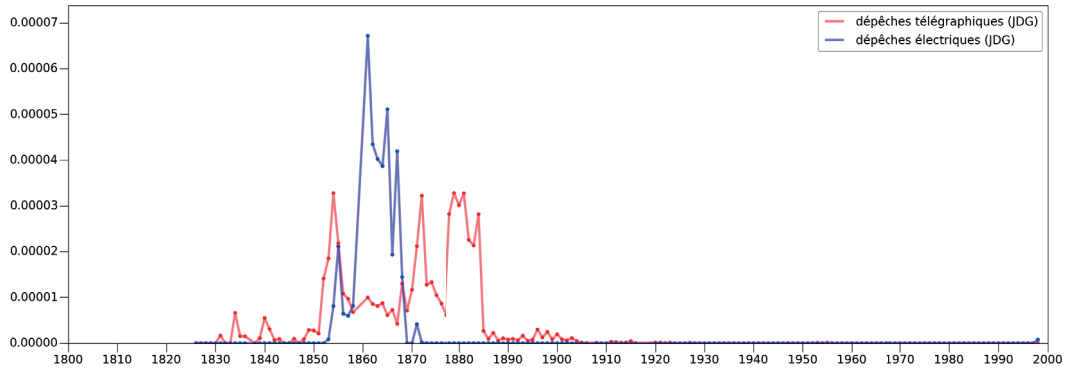


FIGURE 12.58 – Profils fréquentiels de "dépêches télégraphiques" et "dépêches électriques" pour JDG

Nous observons que le 2-gramme "dépêches télégraphiques" subit de fortes baisses de fréquence qui semblent composées par les fréquence de "dépêches électriques". Au sein de JDG le 2-gramme "dépêches électriques" est d'ailleurs celui qui atteint la plus haute fréquence. L'utilisation de ces 2-grammes s'arrête en 1870 pour GDL et 1885 pour JDG pour employer exclusivement le 1-gramme "dépêche" sans plus de besoin de précision.

La prochaine section va nous permettre d'observer à la loupe plusieurs de ces n-grammes et d'autres ainsi que leurs relations les uns par rapport aux autres à l'aide du visualisateur de n-grammes et de la notion de profil fréquentiel.

12.4 Visualisateur de n-grammes

Les chronoclouds nous ont permis de mettre en évidence les profils fréquentiels de n-grammes particuliers. Toutefois, cette visualisation ne donne pas d'informations quant à la relation qu'entretiennent ces n-grammes. Dans cette section nous utilisons le visualisateur de n-grammes afin d'analyser les évolutions fréquentielles de mots intéressants pour l'étude linguistique du corpus tout en cherchant à caractériser les relations qu'entretiennent les n-grammes de différents niveaux n.

Souvent, un mot donné et adéquat au contexte se trouve être trop imprécis. Afin de lui donner de la précision il est d'usage de lui ajouter une suite d'autres mots. Bien entendu, si celle ci est trop longue elle finira par se faire remplacer par un autre mot plus court, mais il est parfois avantageux de garder le mot initial. Prenons l'exemple du mot "Afrique" désignant un vaste continent. On utilisera plutôt le mot "Afrique" en général et de "Afrique du Sud" ou "Afrique du Nord" si l'on souhaite préciser la région. Toutefois, ces notions sont plus complexes dans la langue puisque "Afrique du Sud" est la dénomination d'un pays tandis que la région de l'Afrique du sud est appelée Afrique subsaharienne. Les profils fréquentiels des n-grammes "Afrique", "Afrique du Sud" et "Afrique du Nord" sont présentés dans la Figure 12.59.

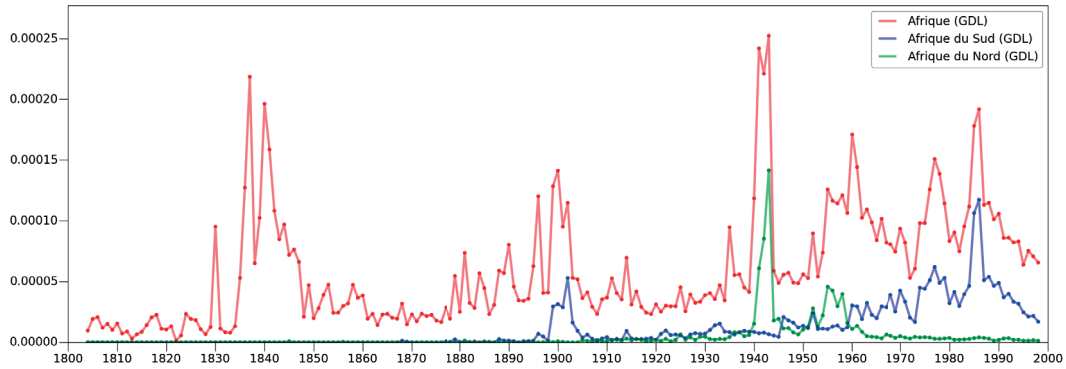


FIGURE 12.59 – Profils fréquentiels de "Afrique", "Afrique du Sud" et "Afrique du Nord" pour GDL

Il est intéressant d'observer les pics fréquentiels de ces n-grammes, leurs fréquences dépendant notamment des événements d'actualité dans ces régions. L'évolution fréquentielle diachronique de chacun de ces n-grammes pris séparément ne relève pas ou peu d'une évolution linguistique, mais plus d'une évolution historique. Toutefois, l'évolution conjointe de ces trois n-grammes peuvent nous apprendre de précieuses informations.

Même si elle est intuitive, nous observons une forme de complémentarité entre ces deux 3-grammes "Afrique du Sud" et "Afrique du Nord", car ils ont des pics de fréquences à des moments différents dans le corpus et cela permet de préciser la source des pics de fréquence du mot général "Afrique".

Un autre exemple est le mot "hockey". Ce sport peut être décliné en plusieurs variantes comme "hockey sur gazon" et "hockey sur glace". Il n'y a que peu de traces de l'utilisation du 3-grammes "hockey sur gazon". Par contre, nous avons détecté que le 3-gramme "hockey sur terre" est utilisé fréquemment. Les profils fréquentiels des n-grammes "hockey", "hockey sur glace", "hockey sur terre" et "hockey sur gazon" sont présentés dans la Figure 12.60.

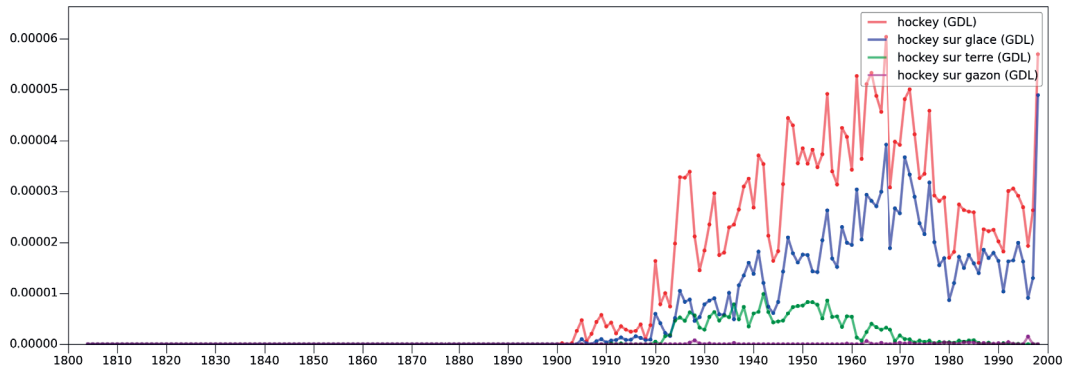


FIGURE 12.60 – Profils fréquentiels de "hockey", "hockey sur glace", "hockey sur terre" et "hockey sur gazon" pour GDL

Nous observons que le 3-gramme "hockey sur gazon" n'était pas utilisé alors qu'il sert aujourd'hui de référence pour ce qui est appelé dans les corpus de JDG et GDL le "hockey sur terre". Toutefois, c'est bien le "hockey sur glace" qui est la plus forte contribution en terme de 3-gramme au profil fréquentiel du 1-gramme "hockey".

Le mot "ministre" désigne une personne par sa fonction, mais au vue de la multitude de ministères différents, il n'est que peu précis si l'on ne cite pas le nom de la personne. Nous illustrons cet exemple canonique en représentant différents n-gramme commençant par le mot ministre sur les Figures 12.61 et 12.62.

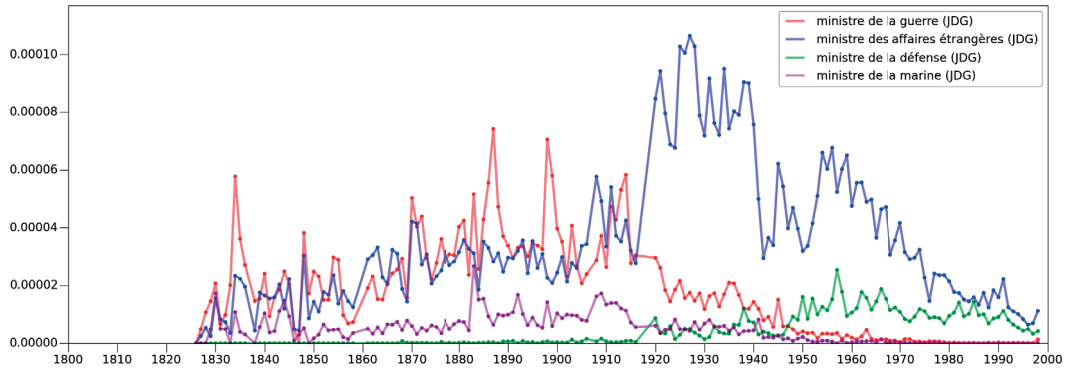


FIGURE 12.61 – Profils fréquentiels de "ministre de la guerre", "ministre des affaires étrangères", "ministre de la défense" et "ministre de la marine" pour JDG

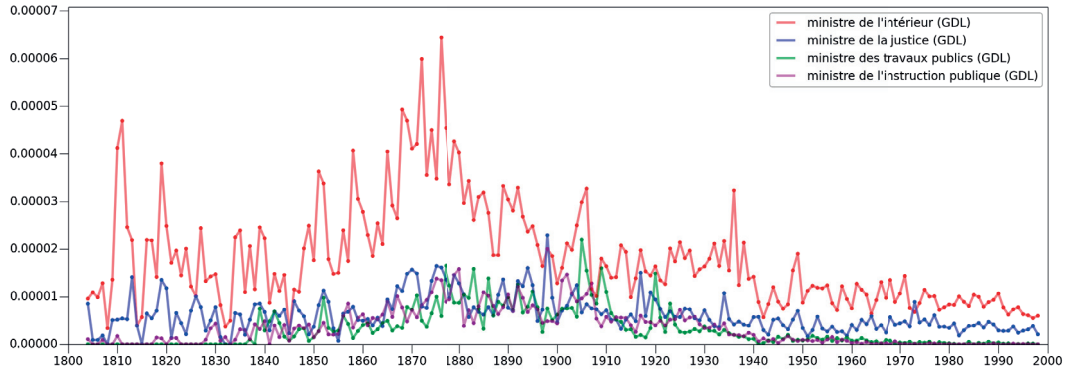


FIGURE 12.62 – Profils fréquentiels de "ministre de l'intérieur", "ministre de la justice", "ministre des travaux publics" et "ministre de l'instruction publique" pour GDL

Nous observons que certains ministères disparaissent ou sont remplacés par d'autres. C'est le cas pour les n-grammes "ministre de la guerre" (remplacé par "ministre de la défense"), "ministre de la marine", "ministre des travaux publics" et "ministre de l'instruction publique".

Outre la précision des personnes, lieux et entités, l'analyse des profils fréquentiels des n-grammes nous permet d'observer les règles d'accord du pluriel. En effet, soit un 2-gramme composé d'un nom et un adjectif comme "homme politique". Nous illustrons les combinaisons d'association de "s" à la fin d'un mot en illustrant les 2-grammes "homme politique", "hommes politique", "homme politiques" et "hommes politiques" dans la Figure 12.63.

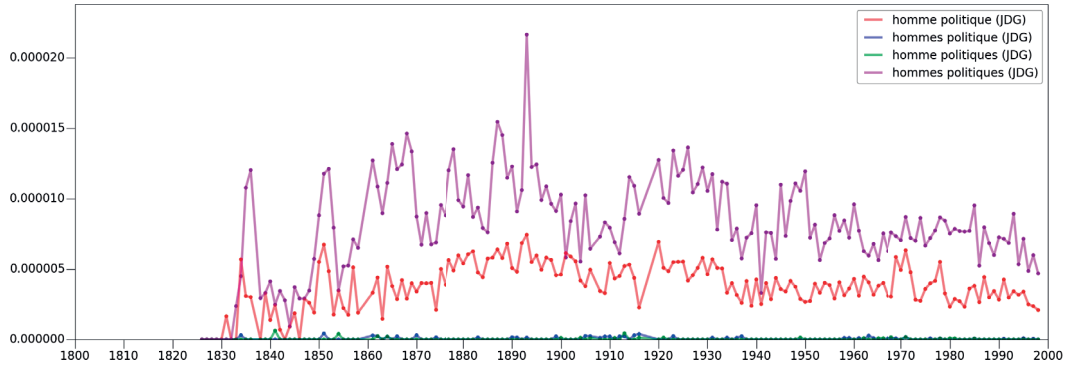


FIGURE 12.63 – Profils fréquentiels de "homme politique", "hommes politique", "homme politiques" et "hommes politiques" pour JDG

Nous observons les formes "autorisées" par la présence en fréquence et celles "interdites" par l'absence en fréquence de celles-ci. Les formes correctes, singulières et plurielles, ont une évolution fréquentielle similaire et une fréquence moyenne comparable. En outre la corrélation entre la forme singulière et plurielle est de 0.59 dans GDL et 0.70 dans JDG.

Il est aussi possible d'observer la règle du genre via le profil fréquentiel. En effet, nous avons déjà observé la présence dans les chronoclouds d'entités précédées par un article. Celui-ci est accordé selon le genre de l'entité. Nous illustrons les combinaisons d'association des déterminants définis "la", "le" et indéfinis "un", "une" en illustrant les 2-grammes "la main", "le main", "un main" et "une main" dans la Figure 12.64.

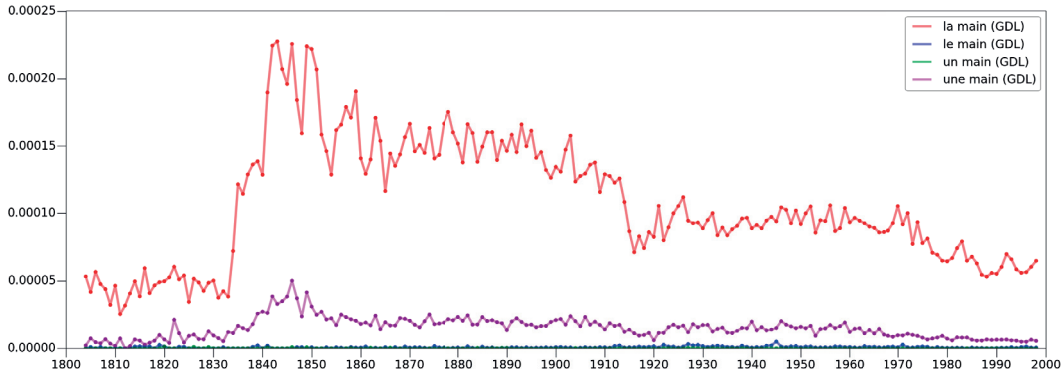


FIGURE 12.64 – Profils fréquentiels de "la main", "le main", "un main" et "une main" pour GDL

Nous observons les formes "autorisées" par la présence en fréquence et celles "interdites" par l'absence en fréquence de celles-ci. Les formes correctes, associant au mot féminin "main" un déterminant féminin comme "la" ou "une", ont une évolution fréquentielle similaire et une fréquence moyenne comparable. En outre la corrélation entre la forme définie et indéfinie est de 0.87 dans GDL et 0.80 dans JDG.

Deux mots peuvent avoir la même forme tout en ayant un genre différent. C'est notamment le cas de "livre" qui au genre masculin désigne un ensemble de pages reliées et au genre féminin désigne la monnaie officielle du Royaume-Uni. Nous illustrons l'association des déterminants définis "la", "le" en illustrant les n-grammes "livre", "le livre" et "la livre" dans la Figure 12.65.

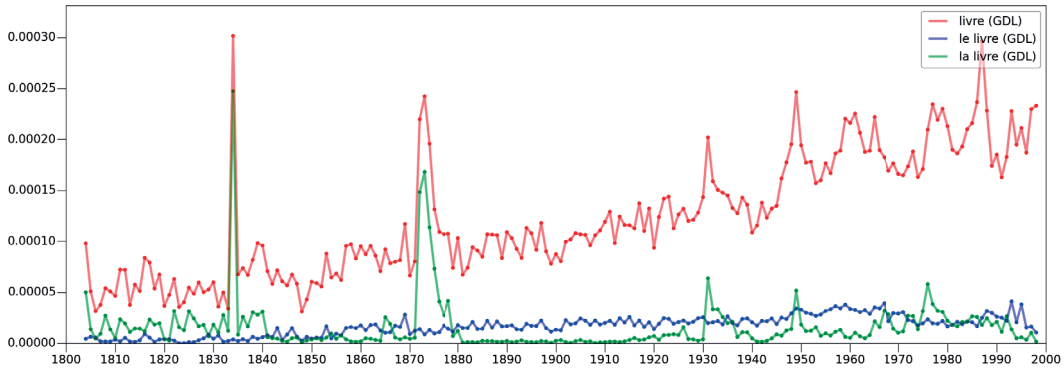


FIGURE 12.65 – Profils fréquentiels de "livre", "le livre" et "la livre" pour GDL

Nous observons que le profil fréquentiel du mot "livre" est en augmentation stable et possède quelques pics de fréquences. Toutefois, il est intéressant de souligner que le profil fréquentiel de "la livre", qui est l'une des décompositions possible du mot "livre", appuie l'hypothèse que les pics fréquentiels du mot "livre" sont uniquement dus à la forme féminine de ce mot et sont donc rattachés au sens de la monnaie plutôt que l'ensemble de pages.

Le profil fréquentiel de "le livre" affiche une faible augmentation tout en restant remarquablement stable dans le corpus. Ainsi ce profil fréquentiel explique la partie stable et en augmentation du profil fréquentiel du mot "livre", partie donc uniquement rattachée au sens de l'ensemble de pages plutôt que la monnaie.

De ces dernières observations, nous remarquons q'il est possible de décomposer un profil fréquentiel d'un mot selon plusieurs sens en étendant le niveau d'étude n à des niveaux plus élevés et en considérant les mots qui suivent ou précédent le mot de référence.

Cela peut se faire d'une façon rétrospective en considérant les mots qui viennent avant le mot de référence (comme dans l'exemple du mot "livre"), mais cela peut se faire de façon prospective également en considérant les mots qui viennent après le mot de référence (comme dans l'exemple du mot "ministre").

Nous retrouvons également la typologie de courbe en peigne dans le corpus, mais dont le cycle est variable via par exemple les profils fréquentiels des 2-grammes "exposition universelle" et "exposition nationale" présentés dans la Figure 12.66.

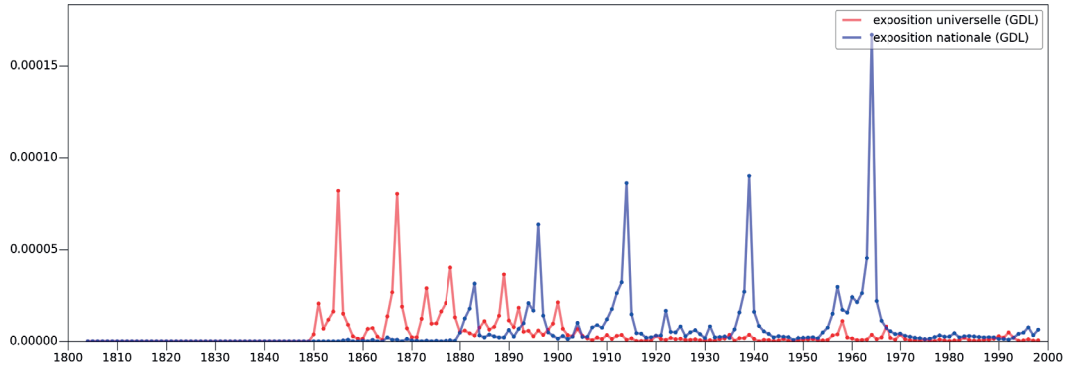


FIGURE 12.66 – Profils fréquentiels de "exposition universelle" et "exposition nationale" pour GDL

Ce type d'exposition a lieu durant certaines années particulières et cela constitue une typologie d'un profil fréquentiel à pics multiples à l'instar du 2-gramme "jeux olympiques". Toutefois, contrairement à "exposition universelle" et "exposition nationale", le cycle de "jeux olympiques" est stable et vaut 4 années (hormis quelques irrégularités tout de même) comme représenté dans la Figure 12.67 où nous montrons les profils fréquentiels de n-grammes commençant par "jeux" et précisant le sens de façon progressive par ajout de mots.

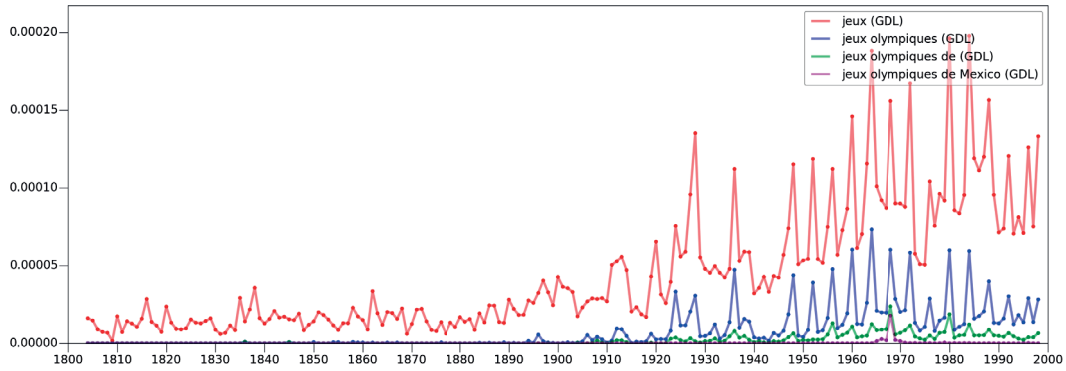


FIGURE 12.67 – Profils fréquentiels de "jeux", "jeux olympiques", "jeux olympiques de" et "jeux olympiques de Mexico" pour GDL

Nous remarquons que la courbe à pics multiples devient une courbe à pic unique dès le niveau $n = 4$. L'ajout de précision par les mots qui suivent "jeux" permet d'isoler un pic et d'expliquer une partie de son profil fréquentiel. En continuant sur cet exemple, nous présentons les profils fréquentiels des 4-grammes "jeux olympiques de Berlin", "jeux olympiques de Londres", "jeux olympiques de Melbourne" et "jeux olympiques de Tokio" dans la Figure 12.68.

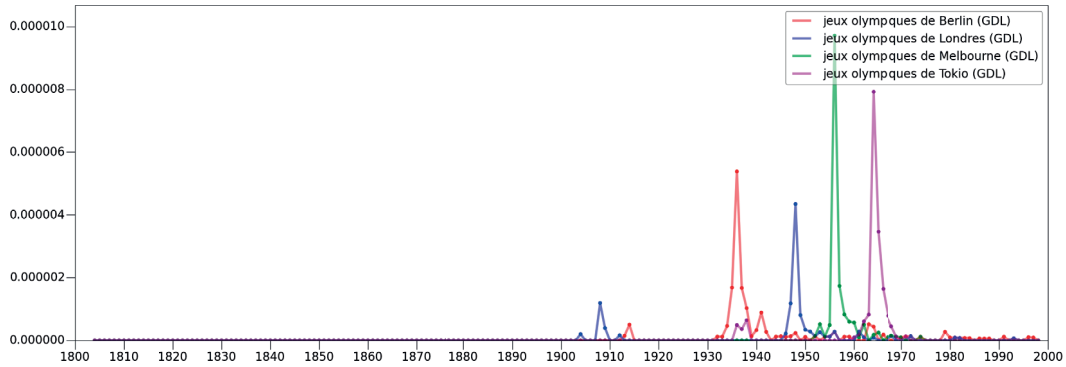


FIGURE 12.68 – Profils fréquentiels de "jeux olympiques de Berlin", "jeux olympiques de Londres", "jeux olympiques de Melbourne" et "jeux olympiques de Tokio" pour GDL

Nous observons que chacun de ces 4-grammes représente un pic de fréquence expliquant l'un des pics de fréquence du 2-gramme "jeux olympiques" ou même du mot "jeux".

Nous pouvons également observer dans le corpus des exemples de variations, apparition et disparition de n-grammes en raison du sens identique de ces deux n-grammes. Par exemple, si l'on considère le 3-grammes "le gouvernement britannique", nous observons la variante "le gouvernement anglais" qui était d'usage dans les années les plus anciennes. Les profils fréquentiels des 3-grammes "le gouvernement britannique" et "le gouvernement anglais" sont présentés dans la Figure 12.69.

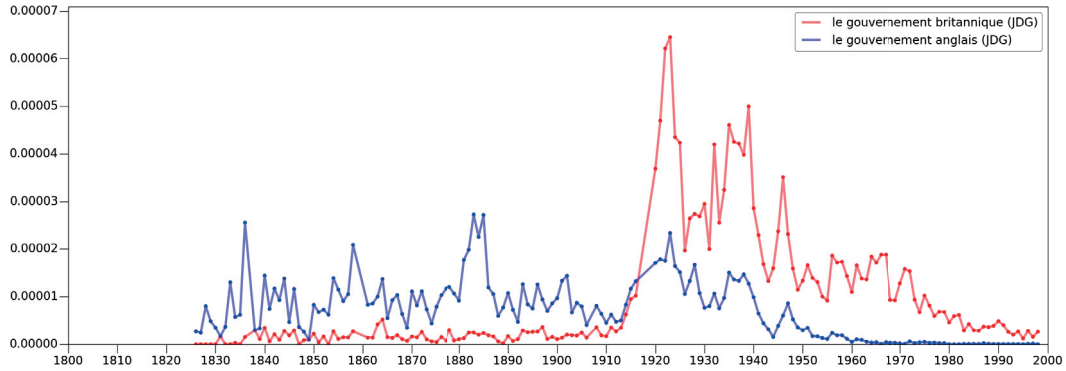


FIGURE 12.69 – Profils fréquentiels de "le gouvernement britannique" et "le gouvernement anglais" pour JDG

Nous observons la disparition progressive (sur environ une cinquantaine d'années) de "le gouvernement anglais" croisant la courbe de "le gouvernement britannique" en 1916. La fréquence de "le gouvernement britannique" augmente brutalement dès son apparition entre 1916 en 1920.

Outre les expressions temporelles et les tournures de phrases qui disparaissent afin de réapparaître sous d'autres formes comme vu dans la section d'analyse des chronoclouds, nous avons aussi remarqué plusieurs exemples de mots relatifs à une expression qui ont tendance à changer de sens avec le temps.

Par exemple, nous considérons le mot "relativement". Ce mot peu être utilisé dans le cadre des expressions "relativement à", "relativement au" et "relativement aux" pour exprimer le sens "à propos" ou "en ce qui concerne". Les profils fréquentiels des n-grammes "relativement", "relativement à", "relativement au" et "relativement aux" sont présentés dans la Figure 12.70.

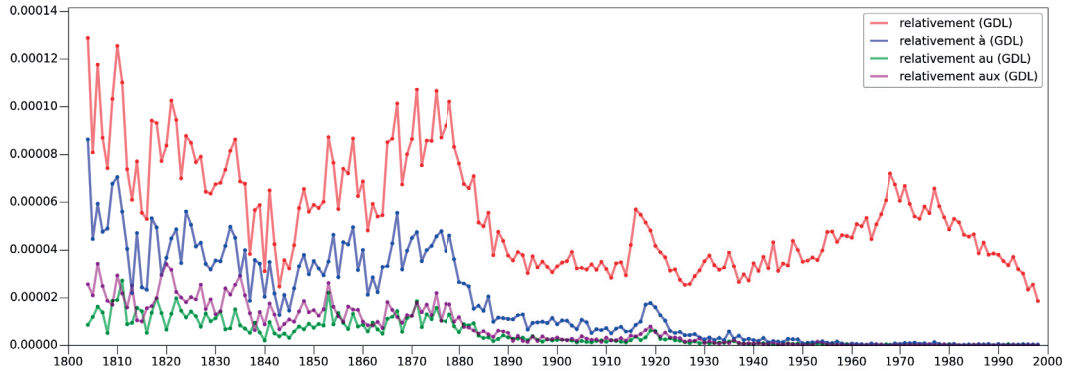


FIGURE 12.70 – Profils fréquentiels de "relativement", "relativement à", "relativement au" et "relativement aux" pour GDL

Nous observons une disparition marquée des expressions "relativement à", "relativement au" et "relativement aux" alors que celles-ci expliquaient majoritairement le profil fréquentiel de "relativement" au cours du XIXième siècle.

Deux questions se posent alors. La première est : par quelles expressions sont remplacées "relativement à", "relativement au" et "relativement aux" afin de maintenir une ou plusieurs expressions ayant le même sens dans le corpus ? Pour répondre au moins partiellement à cette question nous présentons les profils fréquentiels des n-grammes "relativement à la" et "à propos de la" dans la Figure 12.71

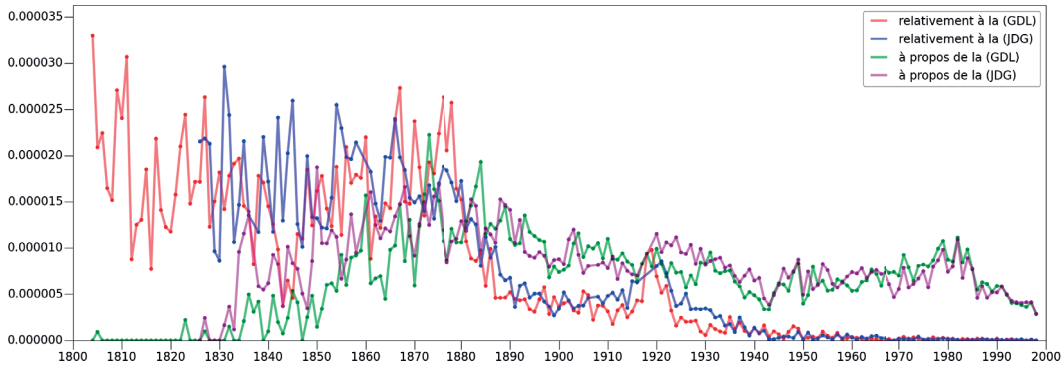


FIGURE 12.71 – Profils fréquentiels de "relativement à la" et "à propos de la"

Nous observons que parallèlement à la baisse du 3-gramme "relativement à la", le 4-gramme "à propos de la" semble augmenter et prendre le relais de cette sémantique.

La deuxième question qui se pose est : quelles sont les expressions utilisant "relativement" au cours du XXe siècle ? En effet, si nous avons constaté que le sens premier de "relativement" a perduré au travers d'autres expressions n'utilisant pas ce mot, "relativement" continue d'être présent dans le corpus et doit donc être explicable.

Nous avons déterminé que l'explication du profil fréquentiel du mot au cours du XXe siècle, au travers par exemple des 2-grammes, est une longue liste composée d'adjectifs suivant "relativement". En effet, le sens de l'utilisation de ce mot a donc changé pour signifier "qui est relatif" dans le sens de "ce qui n'est comme cela que par rapport à autre chose".

Les 2-grammes commençant par "relativement" et formant une colline dans les années les plus récentes sont "relativement peu", "relativement faible", "relativement élevé", "relativement modeste", "relativement facile", "relativement bien", "relativement bas", "relativement important", "relativement calme", "relativement restreint", "relativement bon", "relativement doux", "relativement faibles", "relativement modestes", "relativement moins", "relativement court" et "relativement stable". Les profils fréquentiels des n-grammes "relativement peu", "relativement faible", "relativement élevé" et "relativement modeste" sont présentés dans la Figure 12.72

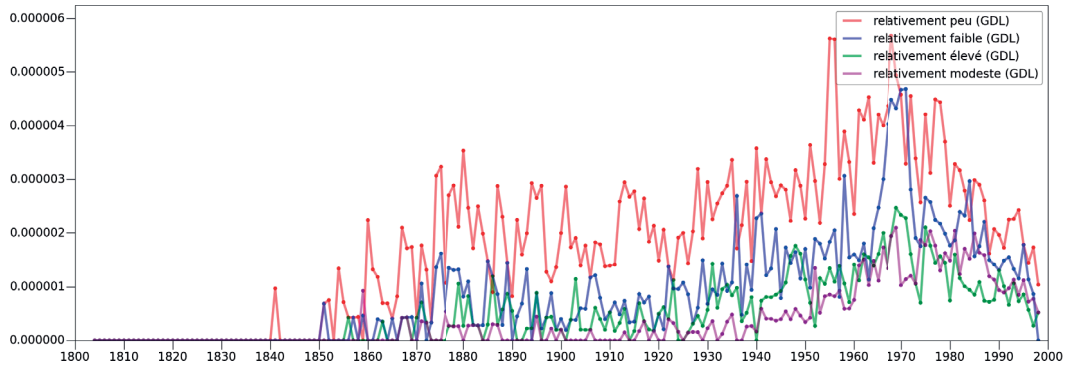


FIGURE 12.72 – Profils fréquentiels de "relativement peu", "relativement faible", "relativement élevé" et "relativement modeste" pour GDL

Nous constatons que ces 2-grammes expliquent en grande partie le profil fréquentiel du mot "relativement" dès le début du XXe siècle.

En outre, nous avons constaté l'apparition de certains anglicismes dans les années les plus récentes. Ces anglicismes dépassent rarement le niveau $n = 2$. Des exemples d'apparitions d'anglicismes sont donnés dans la Figure 12.73.

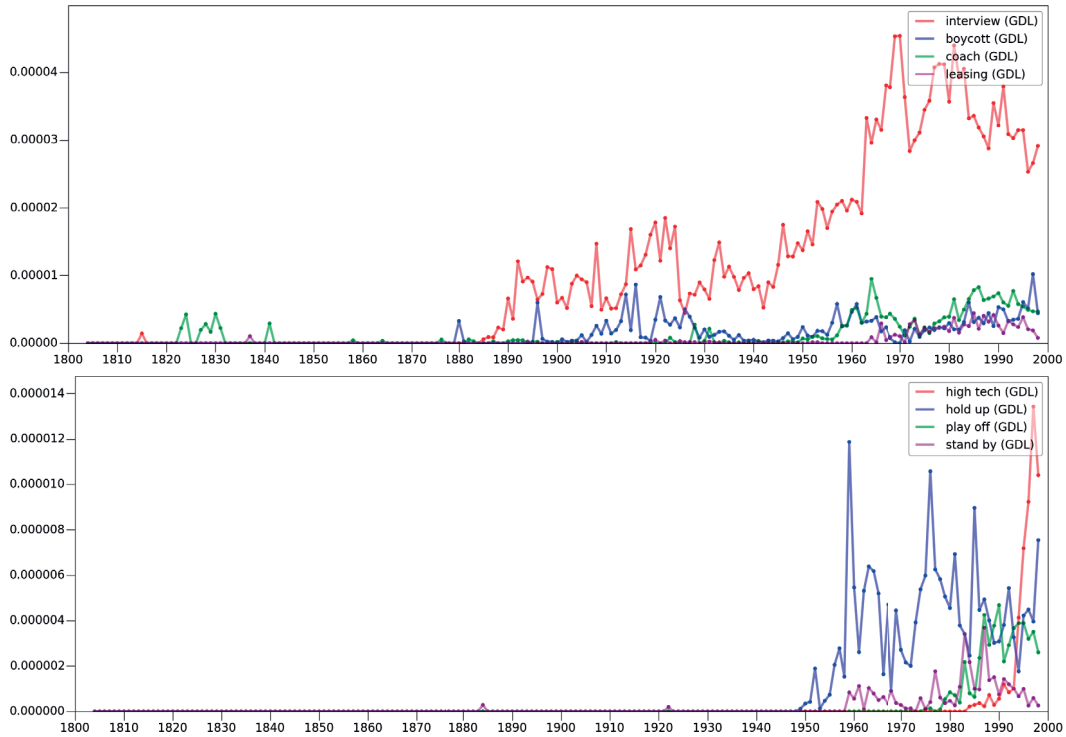


FIGURE 12.73 – Profils fréquentiels de "interview", "boycott", "coach", "leasing", "high-tech", "hold-up", "play-off" et "stand by" pour GDL

13 Synthèse sur les analyses de niveau

Nous avons utilisé les concepts et outils développés dans la première partie de cette thèse afin d'étudier l'espace des n-grammes et de leurs profils fréquentiels selon différents points de vue allant progressivement d'une vision Macro à une vision Micro. Nous avons ensuite réitéré ces mêmes analyses sur les niveaux de n-grammes plus élevés tout en restant systématique sur les conditions et paramètres des méthodes employées.

Nous avons testé la notion de distances nucléaires sur les n-grammes du corpus dans le but de déterminer l'évolution diachronique globale de GDL et JDG tout en réduisant les perturbations non-linguistiques (bourse, horaire de bus, train ou tram, publicités, cinéma, erreurs d'OCR, etc). Nous avons observé un régime linguistique différent dans les années les plus anciennes de 1826-1835 pour JDG et 1804-1835 pour GDL. Toutefois, nos simulations ont montré que la distance nucléaire, comme la distance de Jaccard, est particulièrement sensible dès que l'une de ces années plus lointaines est impliquée en raison de la taille réduite du corpus. Il n'est donc pas possible de conclure à un régime différent de façon significative, mais les analyses du niveaux Micro ont tout de même montré que de nombreux mots changent d'orthographe en raison de la réforme orthographique française de l'année 1835, ce qui correspond à la date exacte du changement de régime observé.

Dès 1835, s'en suit une période relativement stable mais perturbée par deux fois en raison des deux guerres mondiales. Cette période affiche des valeurs similaires sur les deux types de distance qui sont pourtant deux mesures fondamentalement différentes. Il est intéressant de constater que l'impact des deux guerres mondiales est important, mais ne perturbe pas la mesure de la distance sur le long terme, car celle-ci revient ensuite à sa valeur d'avant guerre.

Enfin, un régime particulier est constaté avec la distance de Jaccard sur la période qui va de 1965 jusque 1998. La distance d'une année à l'autre semble être perturbée et varie sans direction claire et les distances séparées par plus d'années augmentent. En effet, nous avons identifié dans cette période plusieurs sources de perturbations non linguistiques. Deux sources dont les effets sont non négligeables sont d'une part les sections de la bourse impliquant un nombre élevé de nombres et de lettres seules et leurs combinaisons ainsi que les grilles de

cinéma qui renforcent les n-grammes plus longs correspondant aux titres de film. La distance nucléaire semble correctement filtrer ces deux sources et montre, au contraire, une stabilité plus grande pour cette dernière période.

La distance nucléaire permet donc de filtrer avec efficacité les perturbations non linguistiques. Toutefois, nous avons déterminé par simulation que celle-ci était encore influencée par les variations de taille du corpus. Deux solutions sont proposées. La première est basée sur l'observation que si la distance nucléaire est impactée par les variations de taille du corpus, elle n'est pas impactée par la taille de vocabulaire du langage dont nous estimons l'évolution, contrairement à la distance de Jaccard. Compte tenu de la remarquable stabilité de l'estimation de cet impact, nous proposons de soustraire les valeurs calculées par simulation sans évolution linguistique et considérant la taille du corpus, aux valeurs calculées directement sur le corpus étudié. Cette solution est basée sur l'hypothèse que ces effets soient en première estimation additifs ce qui n'est pas prouvé.

Cela génère une série de mesures qui, de façon intéressante, mettent en évidence une transition importante en 1835 (réforme orthographique) et une relative stabilité (distance en augmentation légère) jusque 1980 où la distance semble augmenter plus rapidement. Le passage de l'analyse de distance au niveau n supérieur donne des résultats similaires.

La deuxième solution proposée est une combinaison de la notion de noyau résilient non pas avec les distances, mais avec une mesure générale de l'entropie. Il s'agit donc de mesurer la somme des contributions entropiques des n-grammes appartenant uniquement à l'intersection des noyaux résilients des deux journaux. Cette solution à l'avantage de permettre la comparaison de toutes les subdivisions annuelles des deux corpus. Les simulations ont montré que cette mesure est stable et donc fiable uniquement pour les années après 1860.

Il est résulte une valeur d'entropie stable, mais différente pour les deux journaux à $n = 1$ (la différence s'estompe dès $n > 1$) à l'exception de la période de 1940 à 1998 qui exhibe des valeurs d'entropie nucléaire qui diminuent d'années en années. Nous avons donc étudié les variations des contributions entropiques par n-grammes et avons déterminé que cet effet est notamment dû à la diminution constante de la fréquence des n-grammes les plus fréquents. Comme il s'agit d'une étude du noyau résilient, ce n'est donc pas l'apparition de nouveaux n-grammes (qui est également constatée dans cette période), mais bien la hausse de fréquence de n-grammes du noyau résilient, de plus faible fréquence, qui impacte aussi la mesure entropique globale notamment les dix dernières années du corpus.

La visualisation chronocloud nous a permis de mettre en évidence des entités autonomes formées de n mots consécutifs, celles-ci sont identifiables à ce que l'on appelle des expressions multi-mots (multiword expression), expressions fixes et autonomes. Nous avons notamment mis en évidence d'un point de vue plus linguistique les catégories des expressions temporelles et des tournures de phrases qui sont prépondérantes dans le corpus et nous avons constaté que leurs évolutions fréquentielles varient de façon importante dès le début du XXe siècle.

La variante différentielle du chronocloud nous a permis de découvrir des n-grammes dont les évolutions diffèrent dans les deux journaux. Outre les exemples canoniques de lieux et adresses, nous avons trouvé des exemples de changements linguistiques qui consistent en le remplacement d'un n-gramme par un autre pour un même sens donné.

Cela se traduit par une courbe fréquentielle qui tombe relativement brusquement à zéro pour un n-gramme donné pendant qu'un autre n-gramme apparaît et continue la courbe du n-gramme précédent. Le phénomène de remplacement durant lequel les deux n-grammes coexistent a des durées diverses, mais nous avons observé qu'elles durent généralement une dizaine d'années. En outre, nous avons constaté que ces changements connaissent parfois un décalage d'un corpus à l'autre.

Enfin, une analyse purement Micro est effectuée afin de trouver des évolutions particulières de n-grammes intéressants. S'il est hasardeux de tirer des conclusions globales sur des exemples particuliers, nous avons cherché à mettre en évidence la relation entre les différents niveaux n par la même occasion. Nous avons remarqué qu'il est aisé de vérifier les règles grammaticales simples comme les accords du pluriel et du genre.

En outre, de façon intéressante, nous avons observé la résolution de problèmes polysémiques par le passage aux niveaux n plus élevés. L'évolution polysémique diachronique d'un mot peut donc se résoudre, au moins partiellement, par l'étude diachronique des n-grammes incluant le mot donné. En règle générale et sur la base des exemples que nous avons analysés, la polysémie est résolue dès les niveaux $n = 3$ et $n = 4$.

Nous avons trouvé des exemples de mots et expressions dont le sens a changé au cours du temps, modifiant de façon importante une série de n-grammes incluant ce mot. Il est donc possible de détecter les changements sémantiques en considérant les profils fréquentiels d'un mot et n-grammes contenant ce mot. Cette dernière observation nous pousse à analyser plus profondément cette relation dans le chapitre suivant.

14 Analyse multi-échelle

14.1 Distances et Entropie multi-échelle

Dans les chapitres précédents, nous avons analysé les différents niveaux n indépendamment les uns des autres. Dans le cadre de l'analyse diachronique des distances nucléaires et de Jaccard, nous avons observé des résultats globalement similaires entre les premiers niveaux avec néanmoins une perte de qualité dès les niveaux $n > 4$. Dans le cadre de l'analyse de l'entropie nucléaire, nous avons observé des résultats également globalement similaires avec toutefois des différences locales dans les premiers niveaux d'analyse $n < 4$. Malgré la similarité des résultats, certains niveaux d'analyse montrent des variantes subtiles dans les comportements des indices et particulièrement celui de l'entropie nucléaire.

Dans l'optique d'une analyse multi-échelle, nous souhaitons déterminer un indice qui ne dépend pas du niveau n d'analyse. La méthode comporterait alors l'avantage d'être conceptuellement applicable à d'autres types d'éléments comme par exemple les n-grammes de caractères. Nous proposons donc d'agréger les mesures de distances et d'entropie sur toutes les valeurs de n possibles. Une première option possible est un coefficient de pondération qui diminue au fur et à mesure que n augmente. Etant donné les limites d'information (entropie du système) atteintes rapidement avec la longueur n des n-grammes (cf. Figures 12.9 et 12.10), nous proposons un coefficient rapidement dégressif avec n. Une suggestion intuitive serait de pondérer chaque indice par la valeur de l'inverse de n. Etant donné que la somme de ces valeurs ne converge pas, il resterait alors à les normaliser selon le nombre de niveaux n inclus dans l'analyse. L'indice généralisé prend alors la forme suivante :

$$I = \sum_{n=1}^{N} a_n \, I_n = \sum_{n=1}^{N} \frac{\frac{1}{n}}{\sum_{m=1}^{N} \frac{1}{m}} \, I_n = \frac{\sum_{n=1}^{N} \frac{I_n}{n}}{\sum_{m=1}^{N} \frac{1}{m}}$$

où I est l'indice généralisé, I_n l'indice du niveau n des n-grammes, N le niveau n maximal considéré et a_n le coefficient de pondération.

Dans cet exemple de pondération, les coefficients a_n dépendent du niveau maximal N considéré. Ils sont calculés pour $N = 5$ dans la Table 14.1 et pour $N = 9$ dans la Table 14.2.

Niveau (n)	1	2	3	4	5
Pondération (a_n)	44%	22%	15%	11%	9%

TABLE 14.1 – Pondérations a_n en fonction de n avec $N = 5$

Niveau (n)	1	2	3	4	5	6	7	8	9
Pondération (a_n)	35%	18%	12%	9%	7%	6%	5%	4%	4%

TABLE 14.2 – Pondérations a_n en fonction de n avec $N = 9$

Bien que ce modèle de pondération satisfasse à l'objectif d'une agrégation pondérée par des coefficients dégressifs avec le niveau n, il reste dépendant du niveau arbitraire N, soit le total de niveaux pris en compte dans l'indice agrégation. Nous proposons donc de déterminer les coefficients sur la base de la série convergente suivante :

$$\sum_{n=0}^{\infty} r^n = \frac{1}{1-r} \Leftrightarrow \sum_{n=0}^{\infty} (1-r)\, r^n = 1 \qquad r \in \mathbb{R}\,,\ r < 1$$

En considérant la série depuis $n = 1$, nous avons :

$$\sum_{n=1}^{\infty} r^n = \frac{1}{1-r} - 1 = \frac{r}{1-r} = \frac{1}{\frac{1}{r} - 1}$$

La série suivante converge donc vers 1 :

$$\sum_{n=1}^{\infty} a_n = \sum_{n=1}^{\infty} \left(\frac{1}{r} - 1\right) r^n = 1$$

et nous pouvons alors exprimer l'indice généralisé selon l'expression suivante :

$$I = \sum_{n=1}^{\infty} a_n\, I_n = \sum_{n=1}^{\infty} \left(\frac{1}{r} - 1\right) r^n\, I_n$$

Cette forme possède l'avantage d'être indépendante d'un niveau maximal arbitraire N et est plus dégressif que la forme précédente. Ce modèle est donc également applicable aux n-grammes de caractère moyennant le choix d'une valeur de r qui conditionnera la vitesse de convergence de la série des a_n.

Dans cet exemple de pondération, les coefficients a_n sont calculés pour les valeur de r de 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 et 0.99 dans la Table 14.3 en considérant des valeurs de n allant de 1 à 20.

n \ r	0.01	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.99
1	99%	90%	80%	70%	60%	50%	40%	30%	20%	10%	1%
2	1%	9%	16%	21%	24%	25%	24%	21%	16%	9%	1%
3	0%	1%	3%	6%	10%	13%	14%	15%	13%	8%	1%
4	0%	0%	1%	2%	4%	6%	9%	10%	10%	7%	1%
5	0%	0%	0%	1%	2%	3%	5%	7%	8%	7%	1%
6	0%	0%	0%	0%	1%	2%	3%	5%	7%	6%	1%
7	0%	0%	0%	0%	0%	1%	2%	4%	5%	5%	1%
8	0%	0%	0%	0%	0%	0%	1%	2%	4%	5%	1%
9	0%	0%	0%	0%	0%	0%	1%	2%	3%	4%	1%
10	0%	0%	0%	0%	0%	0%	0%	1%	3%	4%	1%
11	0%	0%	0%	0%	0%	0%	0%	1%	2%	3%	1%
12	0%	0%	0%	0%	0%	0%	0%	1%	2%	3%	1%
13	0%	0%	0%	0%	0%	0%	0%	0%	1%	3%	1%
14	0%	0%	0%	0%	0%	0%	0%	0%	1%	3%	1%
15	0%	0%	0%	0%	0%	0%	0%	0%	1%	2%	1%
16	0%	0%	0%	0%	0%	0%	0%	0%	1%	2%	1%
17	0%	0%	0%	0%	0%	0%	0%	0%	1%	2%	1%
18	0%	0%	0%	0%	0%	0%	0%	0%	0%	2%	1%
19	0%	0%	0%	0%	0%	0%	0%	0%	0%	2%	1%
20	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	1%
Total	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**99%**	**88%**	**18%**

TABLE 14.3 – Coefficients a_n calculés pour les valeurs de r de 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 et 0.99 et les valeurs de n allant de 1 à 20

Nous observons que les valeurs extrêmes $r = 0.01$ et $r = 0.99$ ne permettent pas de déterminer une valeur de a_n satisfaisante, car dans le premier cas le poids est exclusivement centré sur le premier niveau et dans la deuxième cas, il est similaire sur quasiment tous les niveaux.

Dans le cas de l'étude des n-grammes de mots sur les corpus de JDG et GDL, il apparaît que la valeur $r = 0.5$ détermine une pondération qui correspond aux objectifs et se comporte d'une manière proche de la mesure de la proportion d'entropie supplémentaire apportée par le niveau n des n-grammes par rapport au niveau précédent $(n-1)$. Cela permet alors de pondérer les niveaux n supérieurs en fonction de l'information supplémentaire apportée au système. Dans le cas de l'application de cette méthode aux n-grammes de caractères, il est préférable de choisir une valeur de r plus élevée afin de tenir compte de n-grammes plus longs. Le choix de $r = 0.5$ nous permet d'écrire l'indice généralisé selon l'expression suivante :

$$I = \sum_{n=1}^{\infty} a_n \, I_n = \sum_{n=1}^{\infty} \left(\frac{1}{r} - 1 \right) r^n \, I_n = \sum_{n=1}^{\infty} \frac{I_n}{2^n}$$

Cette expression constitue notre choix d'agrégation des indices calculés sur les niveaux n.

Bien que la série converge rapidement, il se peut que certaines raisons nous obligent à considérer un niveau maximum N. Par exemple, dans le cas de l'étude des distances nucléaires, le noyau résilient ne contient pas assez de n-grammes dès le niveau $n = 6$. Dans ce cas, il suffit d'appliquer une normalisation sur les premiers coefficients. Ils sont calculés pour $N = 5$ dans la Table 14.4 et pour $N = 9$ dans la Table 14.5.

Niveau	1	2	3	4	5
Pondération	52%	26%	13%	6%	3%

TABLE 14.4 – Pondérations a_n en fonction de n avec $N = 5$

Niveau	1	2	3	4	5	6	7	8	9
Pondération	50%	25%	13%	6%	3%	2%	1%	0%	0%

TABLE 14.5 – Pondérations a_n en fonction de n avec $N = 9$

Nous observons que dans ce cas, la limitation aux 5 premiers niveaux n ne perturbe pas de façon importante les coefficients calculés. Nous optons pour les valeurs de la Table 14.5 dans la cas du calcul des distances de Jaccard et pour les valeurs de la Table 14.4 dans le cas du calcul des distances nucléaires et de l'entropie nucléaire.

Par ailleurs, nous remarquons qu'il est possible de définir un système plus souple en fonction par exemple d'un modèle à plusieurs paramètres. Cela permettrait de considérer également la possibilité de donner plus de poids au 2-grammes et 3-grammes plutôt qu'aux mots. Un exemple de modélisation des pondérations selon une loi log-normale de paramètres μ et σ en fonction du niveau n est donné dans la Table 14.6. La loi s'exprime par la formule suivante :

$$LogN(n;\mu,\sigma) = \frac{1}{n\sigma\sqrt{2\pi}} \exp\left(-\frac{(ln(n)-\mu)^2}{2\sigma^2}\right)$$

n \ par	**LogN(0.4,0.7)**	**LogN(0.6,0.5)**	**LogN(0.7,0.4)**	**LogN(1.2,0.5)**	**LogN(0.8,0.5)**
1	50%	38%	22%	5%	22%
2	27%	38%	50%	24%	39%
3	12%	16%	20%	27%	22%
4	5%	6%	6%	19%	10%
5	3%	2%	2%	12%	4%
6	1%	1%	0%	7%	2%
7	1%	0%	0%	4%	1%
8	0%	0%	0%	2%	0%
9	0%	0%	0%	1%	0%

TABLE 14.6 – Exemple de pondération selon une loi log-normale

Nous observons qu'il est aisé de jouer sur les paramètres μ et σ afin de fournir un vecteur de pondération raisonnable en fonction d'une loi log-normale et du niveau n. De plus, la visualisation de la courbe log-normale associée donne rapidement une idée de la pondération, comme présenté dans la Figure 14.1.

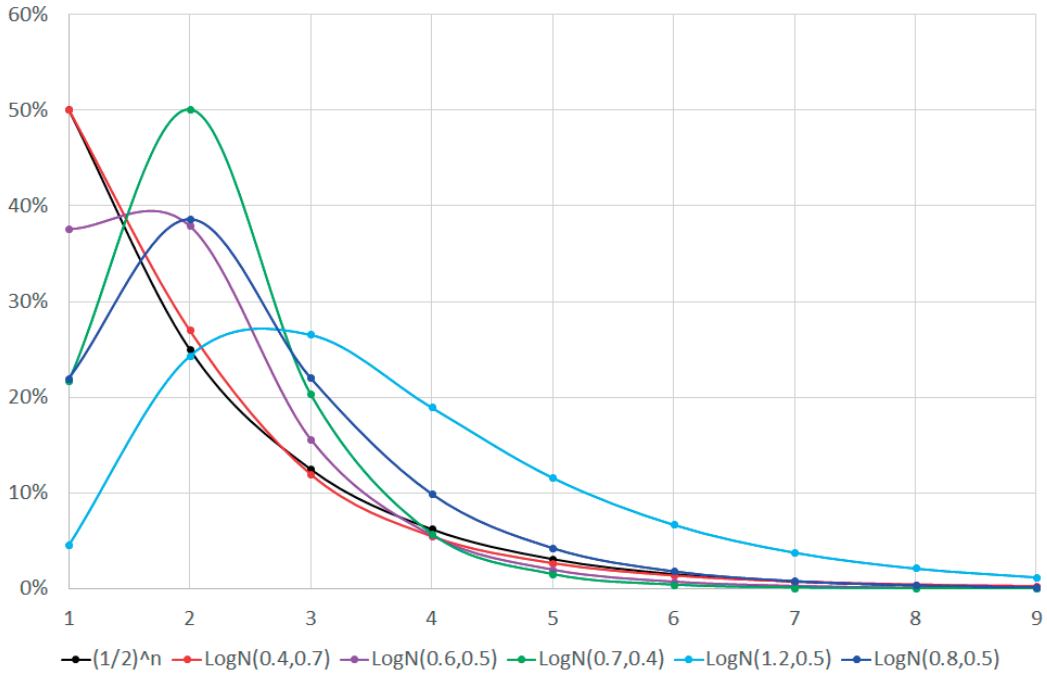


FIGURE 14.1 – Loi log-normale en fonction de n

Par ailleurs, nous observons qu'il est possible de choisir les paramètres μ et σ de façon à retrouver la pondération initiale pour laquelle nous avons opté (cf. les Tables 14.5 et 14.4).

La visualisation graphique des distances de Jaccard et nuclaires multi-échelle selon la pondération choisie est présenté dans la Figure 14.2. Nous y observons des résultats très similaires aux distances de chaque niveau n, car celles-ci sont également similaires à la base. Toutefois, nous observons que les valeurs des distances de Jaccard agrégées ont tendance à être plus élevées en raison du même effet observé au niveau individuel quand le niveau n augmente. Cependant, les périodes que nous avons identifiées précédemment, semblant montrer des régimes d'évolution plus stables, sont conservées par la mesure puisqu'elles sont identifiables pour chaque niveau individuel avec $n < 4$.

Pour la distance nucléaire, les valeurs étant remarquablement stables pour chaque niveau individuel avec $n < 5$, l'indice global a conservé les mêmes comportements. Nous observons donc toujours la stabilité accrue de l'indice de distance nucléaire par rapport à la distance de Jaccard dans les dernières années des corpus. L'indice d'entropie nucléaire a également été agrégé sur la base du même modèle de pondération. Celui-ci est présenté dans la Figure 14.3.

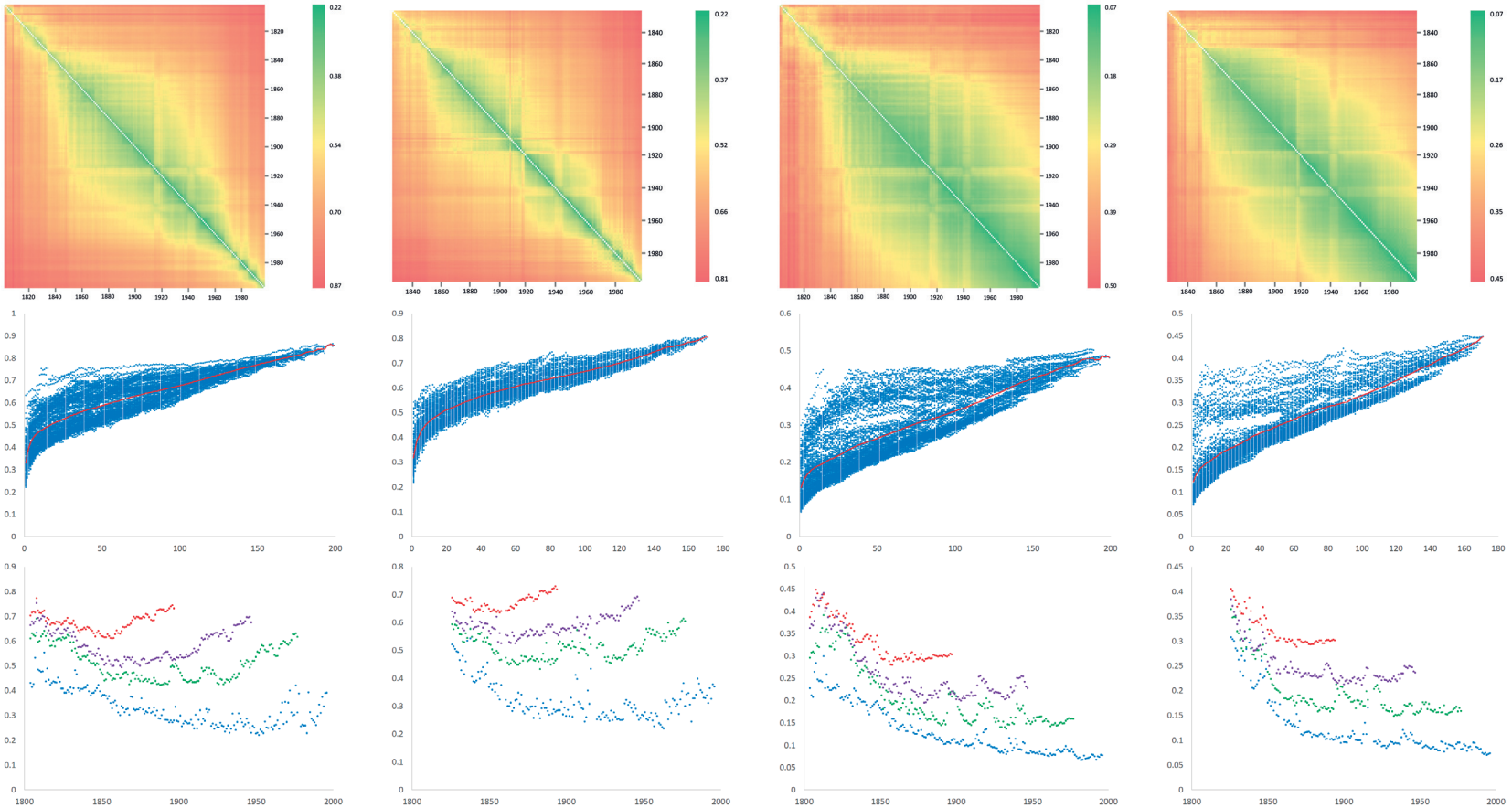


FIGURE 14.2 – **(1)** : Distances de Jaccard sur GDL ; **(2)** : Distances de Jaccard sur JDG ; **(3)** : Distances nucléaires sur GDL ; **(4)** : Distances nucléaires sur JDG ; **Haut** : Heatmap de la matrice des distances ; **Milieu** : Distances (bleu) et moyenne de ces distances (rouge) en fonction du nombre d'année de différence entre les sous-corpus ; **Bas** : Distances entre les années y_i et y_{i+n} avec $n = 1$ (bleu), $n = 20$ (vert), $n = 50$ (violet) and $n = 100$ (rouge)

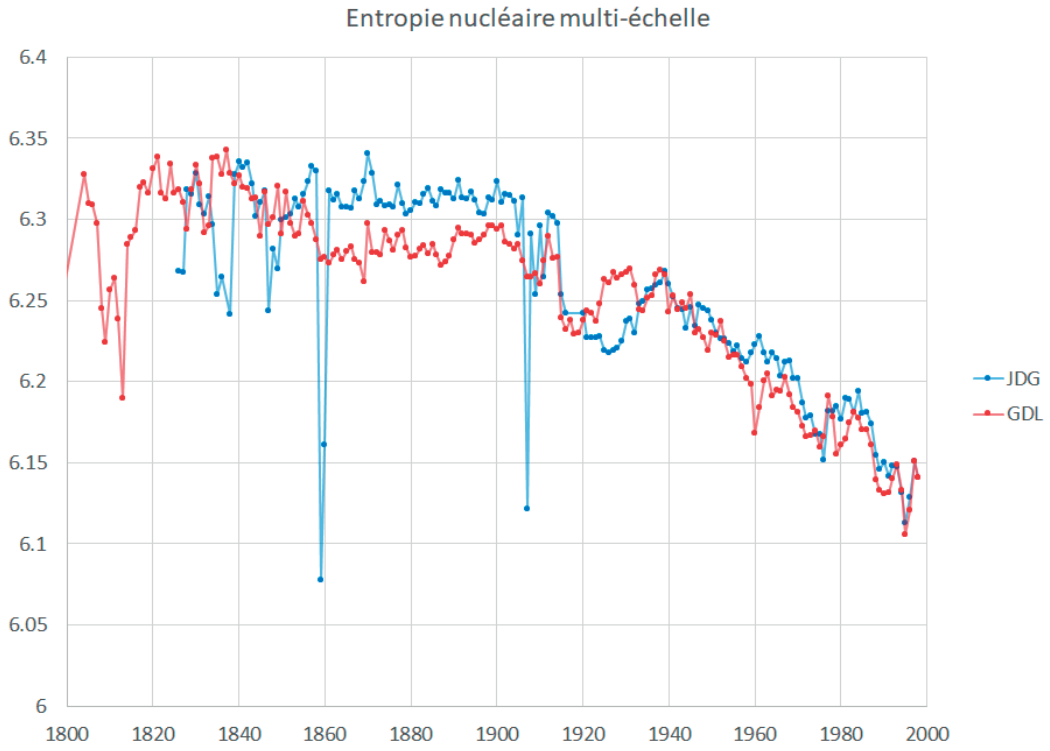


FIGURE 14.3 – Entropie nucléaire multi-échelle

Nous avions observé que la comparaison de l'entropie nucléaire entre les deux corpus présente des différences subtiles pour chaque niveau individuel n, nous souhaitons donc retrouver ces différences dans l'indice de l'entropie nucléaire multi-échelle. Comme selon nos observations aux niveaux n individuels, nous distinguons quatre périodes. La première est constituée des années les plus anciennes soit avant 1860, la seconde va de 1861 à 1906, la troisième va de 1907 à 1940 et la dernière est constituée des années les plus récentes soit après 1941.

L'indice d'entropie nucléaire de la première période est peu stable et les simulations d'effets d'échantillonnage avaient montré que l'indice n'est pas stable pour ces anciennes années. Durant la seconde période, nous avions déjà observé que l'entropie nucléaire de JDG est largement supérieure à celle de GDL pour les 1-grammes uniquement. Selon le niveau individuel n, celle-ci a tendance à augmenter ou diminuer légèrement. Il en résulte que l'entropie nucléaire multi-échelle est particulièrement stable le long de cette période. La troisième période accuse une différence d'entropie nucléaire entre les deux corpus pour les niveaux $n < 4$. En effet, durant cette période l'entropie nucléaire de GDL est légèrement plus élevée que celle de JDG. Il en résulte que cette différence est conservée dans l'indice généralisé. Enfin, la dernière période mettait en évidence une baisse d'entropie selon tous les niveaux n bien qu'elle soit moins prononcée au niveau $n = 1$. Cette baisse se retrouve donc sans équivoque dans l'indice de l'entropie nucléaire multi-échelle.

14.2 Chronocloud multi-échelle

A l'instar des indices du niveau Macro, les visualisations au moyen du chronocloud des niveaux n individuels peuvent être également combinés ensemble en une seule visualisation permettant de représenter l'ensemble des n-grammes de fréquences élevées dans des catégories prédéfinies de résiliences et d'années de fréquence maximale.

Il y a en effet peu de risque d'illisibilité à combiner les n-grammes de niveaux n différents comme la fréquence des n-grammes a tendance à diminuer avec n en raison de la démultiplication des n-grammes d'occurrence unique aux niveaux supérieurs. La proportion de n-grammes représentés dans le chronocloud ne peut donc que diminuer avec le niveau n.

Toutefois, en raison de la séparation classique des espaces entre les mots composant les n-grammes dans la plupart des langages, la proximité des n-grammes représentés au sein d'un chronocloud rend l'identification de la continuité d'un n-gramme difficile.

Il existe plusieurs options, mais nous avons opté pour une solution simple qui requiert de sacrifier l'utilisation initiale de la couleur qui était déjà partiellement représentée par la taille des n-grammes. Le rôle de la couleur était d'effectuer le lien fréquentiel entre les différentes parties du chronocloud.

Dans le cas multi-échelle, nous utilisons la couleur pour représenter le niveau n des n-grammes. Différentes échelles de couleur peuvent être choisies en fonction du design désiré, mais plus les couleurs seront éloignées et plus facile sera la distinction des niveaux n. Pour définir la couleur, nous utilisons le système HSL (Hue Saturation Lightness) aussi appelé TSL en français (Teinte Saturation Luminosité) qui est basé sur la perception des couleurs. Ces trois composantes sont déterminées de la façon suivante :

- La première composante, la teinte, s'exprime comme un entier allant de 0 à 360 et permet de passer de façon continue par les couleurs suivantes : rouge, jaune, vert, bleu, mauve, rose et ensuite le rouge à nouveau.
- La deuxième composante, la saturation, s'exprime comme un nombre réel entre 0 et 1, permet de passer progressivement de la non couleur (gris) vers la couleur.
- La dernière composante, la luminosité, s'exprime comme un nombre réel entre 0 et 1, permet de passer progressivement du noir à la couleur, puis au blanc.

Dans le cas du chronocloud multi-échelle, nous avons opté pour les valeurs suivantes en fonction du niveau n :

- 1-grammes : Bleu, hsl(240, 100%, 30%)
- 2-grammes : Vert, hsl(120, 100%, 30%)
- 3-grammes : Violet, hsl(280, 100%, 30%)
- 4-grammes : Orange, hsl(30, 100%, 30%)
- 5-grammes : Rouge, hsl(0, 100%, 30%)

Les chronoclouds multi-échelle de GDL et JDG sont représentés dans les Figures 14.4 et 14.5.

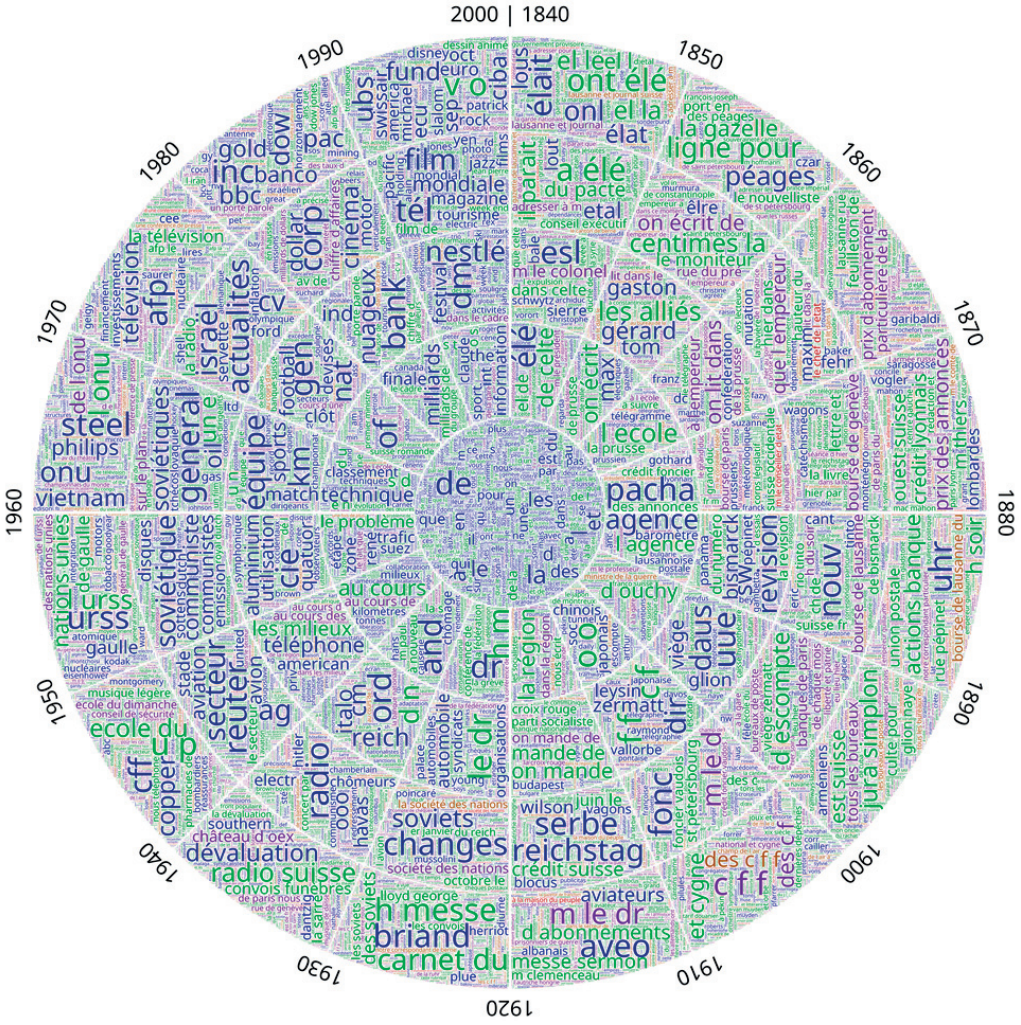


FIGURE 14.4 – Chronocloud multi-échelle pour GDL

La diversité des couleurs permet de visualiser rapidement la répartition des niveaux n par secteur. Le secteur central de résilience $r \geqslant 150$ est essentiellement composé de 1-grammes et 2-grammes. Nous observons également que les secteurs de résilience $50 \leqslant r < 150$ ont une répartition différente selon l'année de fréquence maximale. En effet, les secteurs les plus anciens exhibent plus de n-grammes de niveaux $n > 1$ que les secteurs les plus récents qui exhibent de nombreux mots dont les fréquences sont élevées.

Les observations décrites dans les chapitres précédents pour les niveaux n individuels restent valables pour le chronocloud multi-échelle si ce n'est que la couleur ne permet plus d'observer la variabilité des fréquences entre les secteurs. Cette information est toutefois conservée au sein de chaque secteur. Un n-gramme donné peut être moins visible de par l'addition des n-grammes d'autres niveaux, mais celui-ci peut être retrouvé en utilisant la fonction de zoom.

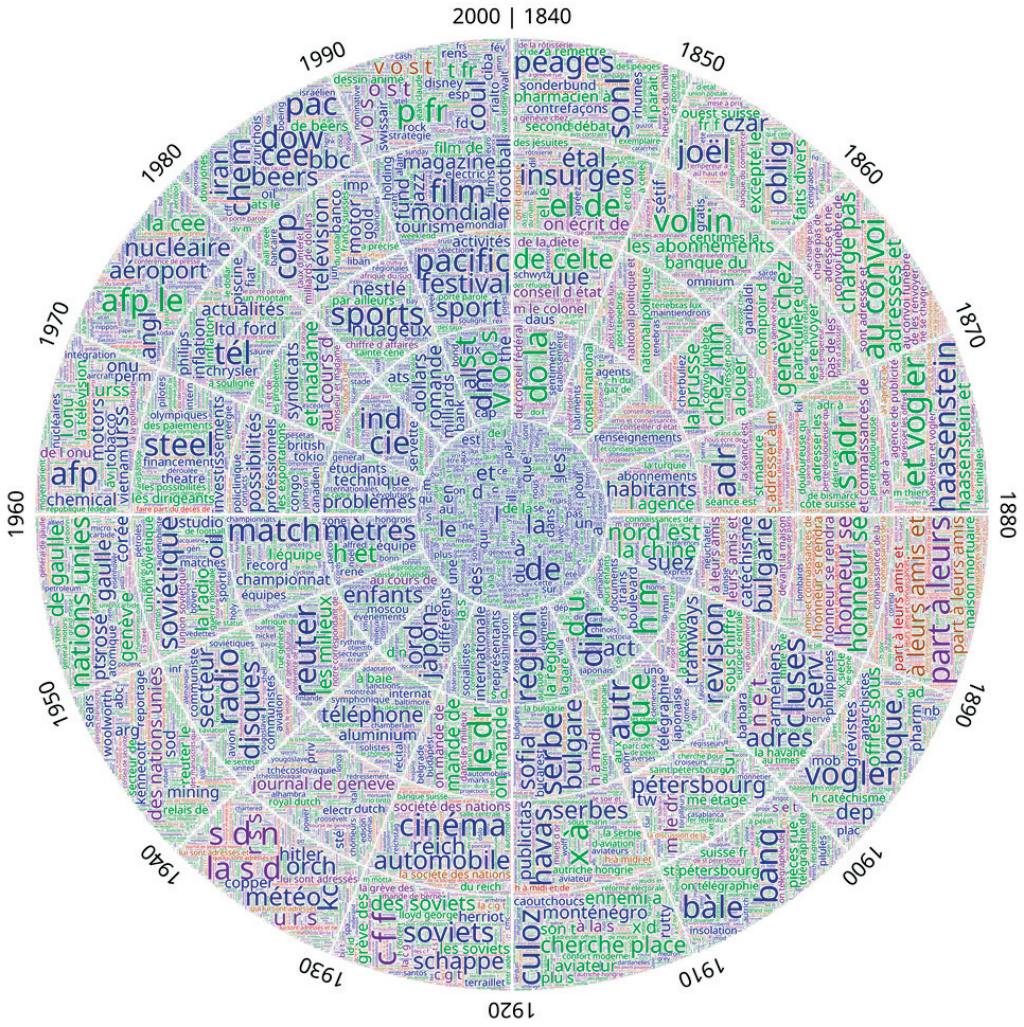


FIGURE 14.5 – Chronocloud multi-échelle pour JDG

Nous observons également que le secteur central de résilience $r \geqslant 150$ est essentiellement composé de 1-grammes et 2-grammes. Nous observons aussi que les secteurs les plus anciens exhibent plus de n-grammes de niveaux $n > 1$ (particulièrement des 2-grammes et 3-grammes) que les secteurs les plus récents qui exhibent plutôt des mots. Cet effet est prononcé dans les cas du secteur correspondant aux années 1880 à 1890 du journal de JDG. Les n-grammes les plus fréquents observés dans ce secteur sont associés aux annonces type de décès que nous avions déjà identifiés dans les chronocloud de niveau n individuel.

Nous avons donc réuni les n-grammes de tous niveaux en une seule visualisation en sacrifiant une partie de l'information fréquentielle. Cette visualisation possède une densité informationnelle importante et d'autant plus en considérant la possibilité de zoomer dans le cas d'un calcul d'une image de haute qualité munie d'une technologie Deep Zoom.

14.3 Décomposition multi-échelle des profils fréquentiels

Nous avons précédemment développé une étude des profils fréquentiels des n-grammes sur les niveaux n individuels du corpus de GDL et JDG avec pour objectif de cibler autant que possible les évolutions qui relèvent du changement linguistique plutôt que culturel ou historique. Cependant, une dimension généralement peu explorée dans ce type d'analyse est la relation entre le niveau n des n-grammes et le genre de phénomènes étudiés.

Intuitivement, l'analyse basée sur les 1-grammes capture les caractéristiques globales du changement de vocabulaire et les tendances relevant de l'évolution des sujets traités par les deux journaux. L'analyse des niveaux intermédiaires ($1 < n < 6$) permet la capture d'expressions multi-mots (multiword expressions) qui ont un comportement autonome dans le langage ("Prix d'affaire", "maison de retraite", "ministre des affaires étrangères", etc) ainsi que des expressions syntaxiques et tournures de phrase ("il y a", "c'est à dire", "quoi qu'il en soit", ect). Enfin, les niveaux supérieurs ($n > 5$) correspondent plutôt à des répétitions volontaires comme des annonces officielles (naissance officielle, annonce d'assemblée générale, etc) ou simplement du plagiat de type "copier-coller".

Le visualisateur de n-grammes a l'avantage de permettre la visualisation et la comparaison de profils fréquentiels de n-grammes de niveaux différents. Nous avons déjà vu que certains profils fréquentiels entretiennent des relations particulières permettant d'identifier des changements sémantiques (Par exemple, le mot "relativement" cf. Figure 12.70, 12.71 et 12.72). Cet exemple, montre qu'il est possible de repérer une évolution sémantique sur la base de la relation d'un mot avec les 2-grammes qui contiennent ce mot.

Dans le chapitre précédent, nous avons étudié la relation entre les n-grammes qui n'ont pas de chaînes de mots en commun et avons conclu que les comparaisons de profils fréquentiels permettent d'extraire de l'information sémantique sur les n-grammes. Dans ce chapitre, nous étudions la relation entre le profil fréquentiel d'un n-grammes avec ceux d'autres n-grammes partageant une suite commune de mots.

Nous présentons dans ce chapitre une analyse linguistique de corpus basé sur la relation canonique permettant de décomposer un profil fréquentiel d'un n-gramme en la somme des profils fréquentiels des (n+1)-grammes le contenant.

Afin de visualiser d'une façon naïve une partie du contenu des ensembles de n-grammes de différents niveaux, nous présentons une liste des n-grammes les plus fréquents du corpus de JDG dans la Table 14.7 pour tous les niveaux $n < 6$.

1-grammes		2-grammes		3-grammes		4-grammes		5-grammes	
de	0.045391	de la	0.007004	n est pas	0.000404	ce n est pas	9.62E-05	il n y a pas	4.82E-05
l	0.019357	à l	0.002178	n a pas	0.000271	le conseil d etat	0.000112	ans v o s t	2.01E-05
à	0.019061	à la	0.002689	il y a	0.000416	il n y a	0.000116	sous la présidence de m	2.64E-05
la	0.025518	de l	0.004356	que l on	0.00025	en ce qui concerne	6.99E-05	en ce qui concerne les	2.04E-05
le	0.019029	tous les	0.000748	il n y	0.000188	c est à dire	9.11E-05	o s t fr all	3.68E-05
les	0.015949	a été	0.001126	et de la	0.000213	du journal de genève	6.35E-05	amis et connaissances de la	2.05E-05
d	0.014622	que les	0.000941	le conseil d	0.000168	il n est pas	6.45E-05	de la s d n	2.41E-05
des	0.012364	que l	0.00084	de la société	0.000177	du conseil d etat	6.58E-05	de la ville de genève	2.69E-05
et	0.018634	que le	0.001089	de la ville	0.000171	v o s t	5.78E-05	à l ordre du jour	2.10E-05
a	0.010566	que la	0.000895	conseil d etat	0.000221	o s t fr	4.39E-05	à 12 h et de	2.10E-05
que	0.008757	c est	0.001494	journal de genève	0.000136	au point de vue	4.72E-05	v o s t fr	4.35E-05
qui	0.006942	il a	0.000663	c est à	0.000153	s t fr all	4.42E-05	qu il y a de	1.35E-05
il	0.007691	il y	0.000609	et de l	0.000163	la ville de genève	4.38E-05	département de justice et police	1.23E-05
est	0.007536	il est	0.000681	ce n est	0.000164	l ordre du jour	5.14E-05	ans p fr 14 00	1.54E-05
en	0.009863	et l	0.00059	ce qu il	0.000119	de plus en plus	5.23E-05	charge pas de les renvoyer	1.11E-05
m	0.00614	et les	0.000787	le conseil fédéral	0.00012	de la ville de	5.52E-05	leurs amis et connaissances de	1.87E-05
au	0.005822	et de	0.001447	de la commission	0.000143	n y a pas	5.44E-05	que le conseil d etat	1.48E-05
s	0.006657	et le	0.000723	de la suisse	0.000149	n a pas été	4.19E-05	qui lui sont adressés et	1.11E-05
pour	0.006162	et la	0.00071	n y a	0.000136	à 20 h 30	5.03E-05	font part à leurs amis	1.26E-05
une	0.007604	sur les	0.000638	qu il a	0.000167	qu il y a	6.08E-05	il y a quelques jours	1.37E-05
dans	0.006433	sur le	0.00072	qu il avait	7.02E-05	du conseil d administration	2.15E-05	sous la direction de m	1.39E-05
un	0.008404	sur la	0.000719	qu il s	6.61E-05	du chemin de fer	2.58E-05	en ce qui concerne le	1.21E-05
par	0.005482	s est	0.000755	qu il ne	0.000112	conseil d etat a	2.06E-05	en ce qui concerne la	1.71E-05
n	0.004611	pour les	0.000545	qu il y	0.000114	ans p fr 14	2.26E-05	et ne se charge pas	1.17E-05
du	0.010359	pour la	0.000621	qu il est	9.97E-05	ans v o s	2.16E-05	et connaissances de la perte	1.73E-05
tous	0.001193	y a	0.000632	rue de la	8.44E-05	des chemins de fer	3.93E-05	à leurs amis et connaissances	1.83E-05
t	0.001828	dans l	0.000595	millions de francs	6.93E-05	journal de geneve du	4.04E-05	12 h et de 14	1.90E-05
être	0.00152	dans les	0.000892	où l on	7.40E-05	leurs amis et connaissances	2.25E-05	14 00 16 00 18	1.84E-05
même	0.001334	dans le	0.001095	du grand conseil	6.45E-05	qui n est pas	2.70E-05	part à leurs amis et	1.91E-05
où	0.001075	dans la	0.001078	du conseil d	0.000107	c est ainsi que	2.64E-05	m le conseiller d etat	1.48E-05
ou	0.001733	ont été	0.000551	du journal de	9.03E-05	il s agit de	3.48E-05	ce qu il y a	1.63E-05
conseil	0.001371	d un	0.001424	point de vue	9.87E-05	il s agit d	2.73E-05	au point de vue de	1.23E-05
était	0.001221	d une	0.001479	ans p fr	6.52E-05	ministre des affaires étrangères	3.50E-05	00 16 00 18 00	1.34E-05
comme	0.0015	l on	0.000598	projet de loi	8.06E-05	il y a quelques	3.60E-05	00 18 00 20 00	1.34E-05
c	0.002753	de son	0.000522	journal de geneve	7.86E-05	il y a des	2.21E-05	00 20 00 22 00	1.28E-05
e	0.002281	de genève	0.000644	a eu lieu	8.17E-05	il y a eu	3.37E-05	quoi qu il en soit	1.59E-05
h	0.002325	de m	0.00088	a t il	7.58E-05	il y a un	2.28E-05	s t fr all 18	1.37E-05
aux	0.002375	de ce	0.000508	que le gouvernement	8.03E-05	il est vrai que	2.13E-05	sont adressés et ne se	1.10E-05
lui	0.001442	par les	0.000522	que c est	6.46E-05	l honneur se rendra	3.25E-05	le conseil d etat a	1.75E-05
i	0.002568	par le	0.000714	qui s est	7.73E-05	de la société des	3.19E-05	le ministre des affaires étrangères	1.15E-05
1	0.003931	par la	0.000614	qui a été	7.22E-05	à la fin de	3.34E-05	le président de la république	1.17E-05
bien	0.001312	n est	0.000743	c est que	7.09E-05	il n a pas	3.43E-05	la fin de l année	1.21E-05
3	0.002609	n a	0.000653	c est une	7.48E-05	il n y avait	2.42E-05	fr 14 00 16 00	1.37E-05
été	0.002497	qu il	0.0017	c est un	8.78E-05	à la suite de	3.13E-05	président du conseil d etat	1.21E-05
5	0.002853	qu ils	0.000419	c est le	9.48E-05	il ne faut pas	3.14E-05	l u r s s	1.44E-05
sur	0.003877	par l	0.000391	c est la	8.40E-05	à 12 h et	2.47E-05	p fr 14 00 16	1.64E-05
ce	0.003978	à une	0.000381	20 h 30	0.000111	14 00 16 00	2.87E-05	de l hôtel de ville	1.66E-05
25	0.00145	aujourd hui	0.000457	la ville de	0.000106	sous la présidence de	2.98E-05	de la société des nations	1.28E-05
20	0.001498	qu elle	0.000477	président de la	0.00011	sous la direction de	2.26E-05	de la perte douloureuse qu	1.15E-05
se	0.003921	à genève	0.000349	de tous les	0.000108	en la personne de	2.47E-05	par le conseil d etat	1.58E-05

TABLE 14.7 – Les cinquante n-grammes les plus fréquents du corpus de JDG pour $n < 6$

Nous observons que les 1-grammes les plus fréquents correspondent souvent à des articles, pronoms, adverbes et prépositions ("de", "l", "à", "la", "le"). De la même façon, les 2-grammes les plus fréquents correspondent pour la plupart à des combinaisons d'articles, pronoms, adverbes et prépositions ("de la", "à l", "à la", "de l", "tous les"). Il est intéressant d'observer parmi les 3-grammes les plus fréquents qu'en plus de ces combinaisons basiques ("n'est pas", "n'a pas", "il y a", "que l'on", "il n'y"), apparaissent des expressions multi-mots ("Conseil d'état", "le conseil fédéral", "du grand conseil", "point de vue" et "projet de loi").

Ces expressions multi-mots sont encore plus nombreuses parmi les 4-grammes les plus fréquents ("la ville de genève", "l'ordre du jour", "du conseil d'administration", "des chemins de fer", "ministre des affaires étrangères"). Ces apparitions d'expressions multi-mots continuent également pour les 5-grammes les plus fréquents si ce n'est qu'apparaissent également des expressions figées correspondant à des phrases types d'annonces officielles répétées volontairement dans certaines section du journal (par exemple l'expression "... font part à leurs amis et connaissances de la perte douloureuse qu'éprouve ...").

La sémantique d'une expression multi-mots se trouve dans la combinaison exacte des mots qui la composent. Pour la plupart de ces expressions, les mots individuels ne donnent que peu d'indications quand au sens exact de ces entités. Prenons, par exemple, le mot "maison" qui a généralement le sens d'un bâtiment que l'on habite. Si nous faisons suivre ce mot par certaines combinaisons particulières nous pouvons obtenir "maison blanche", "maison mortuaire", "maison de commerce" ou "maison de paroisse". Chacune de ces expressions multi-mots n'ont que peu de sémantique en commun vis-à-vis du mot "maison". Il faut observer qu'il ne suffit pas qu'une suite de mots soit fréquente pour qu'elle ait un sens bien précis puisque l'ajout du mot "de" derrière "maison" ne permet pas de résoudre le sens de façon précise. Il semble donc que l'ajout progressif de mots à des fins de précision permet d'écarter successivement diverses sémantiques jusqu'à tomber sur un sens précis qui ne souffre d'aucune ambiguïté. Dès lors, l'expression peut se figer et se comporter de façon autonome.

Prenons l'exemple de "maison blanche". Dans la logique de la langue française, ce 2-gramme signifierait une maison au sens du 1-gramme de la catégorie des noms communs suivi de l'adjectif spécifiant la couleur blanche de cette dernière. Toutefois, l'utilisation de ce 2-gramme possède aussi un autre sens qui ne peut pas se déduire de celui de ses composantes. Il désigne le lieu où siège et vit le président des Etats-Unis. Bien entendu, le lien existe, car ce lieu correspond à un bâtiment de couleur blanche. Mais dans ce cas, le mot "blanche" n'est plus un adjectif, mais complète le mot maison pour créer une entité autonome agissant comme un nom commun. Dans ce cas particulier, nous notons que "maison blanche", de son premier sens, peut être précédé par les articles "une" ou "la". En revanche, s'agissant du deuxième sens, le 2-gramme n'est jamais précédé de "une", car cela contredirait la logique de précision voulant qu'il n'existe qu'une seule maison blanche qui correspond au siège du président des Etats-Unis. Les profils fréquentiels de "maison blanche", "la maison blanche" et "une maison blanche" présentés dans la Figure 14.6 ne laissent que peu de doutes sur le fait que, dans ce corpus, "maison blanche" signifie presque exclusivement la demeure du président des Etats-Unis notamment en observant que "une maison blanche" n'est quasiment pas présent ni dans JDG ni dans GDL et que "maison blanche" semble indissociable de l'article "la".

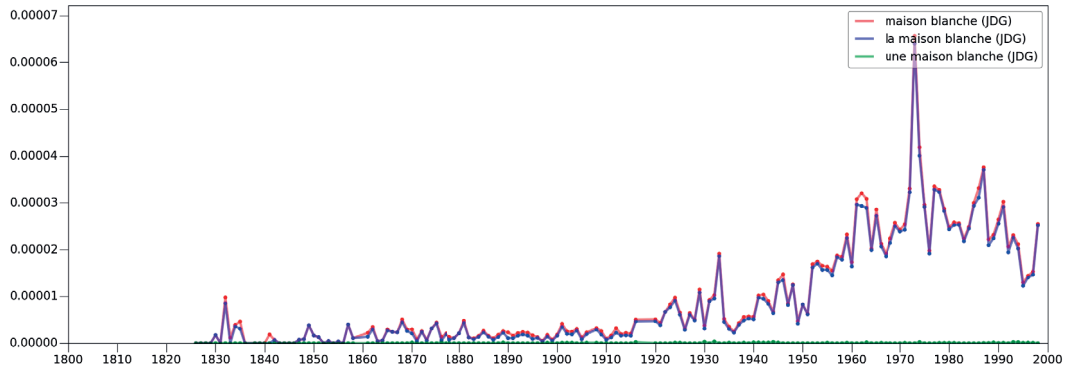


FIGURE 14.6 – Profils fréquentiels de "maison blanche", "la maison blanche" et "une maison blanche" pour JDG

Afin de se donner une idée des différents sens que peuvent engendrer les n-grammes commençant par le mot maison nous représentons les vingt n-grammes les plus fréquents commençant par "maison" pour JDG dans la Table 14.8 pour $1 < n < 6$.

niveau 2		niveau 3		niveau 4		niveau 5	
2-grammes	Frequences	3-grammes	Frequences	4-grammes	Frequences	5-grammes	Frequences
maison du	1.06E-05	maison de commerce	6.01E-06	maison de commerce de	1.49E-06	maison de commerce de la	7.23E-07
maison bourgeoise	1.77E-06	maison d habitation	5.95E-06	maison des arts du	1.42E-06	maison mortuaire rue de la	5.32E-07
maison n	2.27E-06	maison mortuaire rue	5.81E-06	maison mortuaire rue de	1.41E-06	maison de la lyre d	6.01E-07
maison il	1.61E-06	maison de la	1.01E-05	maison mortuaire rue du	1.28E-06	maison de paroisse de la	2.03E-07
maison se	1.59E-06	maison de ville	9.31E-07	maison communale de plainpalais	1.28E-06	maison de paroisse de plainpalais	3.28E-07
maison meublée	1.94E-06	maison de force	7.57E-07	maison de la poste	1.76E-06	maison de paroisse de st	1.08E-07
maison en	3.35E-06	maison de l	7.24E-07	maison de paroisse 11	5.02E-07	maison de paroisse bourg de	1.05E-07
maison et	7.05E-06	maison de gros	1.22E-06	maison de la radio	5.20E-07	maison de paroisse rue dassier	1.20E-07
maison soussignée	1.59E-06	maison de détention	1.03E-06	maison de la jeunesse	6.86E-07	maison de paroisse 11 h	2.36E-07
maison s	3.09E-06	maison de l	2.27E-06	maison de la rue	7.25E-07	maison de paroisse 11 rue	1.53E-07
maison où	3.34E-06	maison de jeu	1.04E-06	maison de la suisse	6.80E-07	maison de paroisse 10 h	1.54E-07
maison blanche	9.08E-06	maison de banque	2.37E-06	maison de la place	7.13E-07	maison de paroisse des eaux	2.69E-07
maison à	4.59E-06	maison de 10	6.93E-07	maison de la lyre	6.59E-07	maison de paroisse 9 h	2.57E-07
maison des	1.11E-05	maison de correction	6.95E-07	maison de 10 pièces	5.84E-07	maison de 1 er ordre	2.36E-07
maison a	3.15E-06	maison de premier	6.84E-07	maison de premier ordre	6.31E-07	maison de gros de la	2.52E-07
maison située	1.48E-06	maison de quartier	1.02E-06	maison de quartier de	5.77E-07	maison de la radio 66	1.43E-07
maison avec	1.53E-06	maison de m	1.42E-06	maison de 12 pièces	5.20E-07	maison de la jeunesse 3	1.63E-07
maison dans	2.26E-06	maison de retraite	7.52E-07	maison de commerce ou	5.42E-07	maison de la poste au	1.34E-07
maison neuve	2.32E-06	maison de savoie	1.23E-06	maison de campagne meublée	8.44E-07	maison de la poste blanc	1.96E-07
maison d	1.80E-05	maison de santé	1.41E-06	maison de campagne de	6.71E-07	maison de la poste avis	1.17E-07

TABLE 14.8 – Les vingt n-grammes les plus fréquents commençant par "maison" pour JDG

Nous observons dans ce tableau qu'apparaissent un grand nombre de expressions multi-mots commençant par le mot "maison". En analysant le profil fréquentiel de certains de ces n-grammes, nous observons qu'ils ont une grande disparité temporelle. Nous présentons les profils fréquentiels des n-grammes "maison de paroisse", "maison de commerce", "maison blanche" et "maison de quartier" pour JDG dans la Figure 14.7.

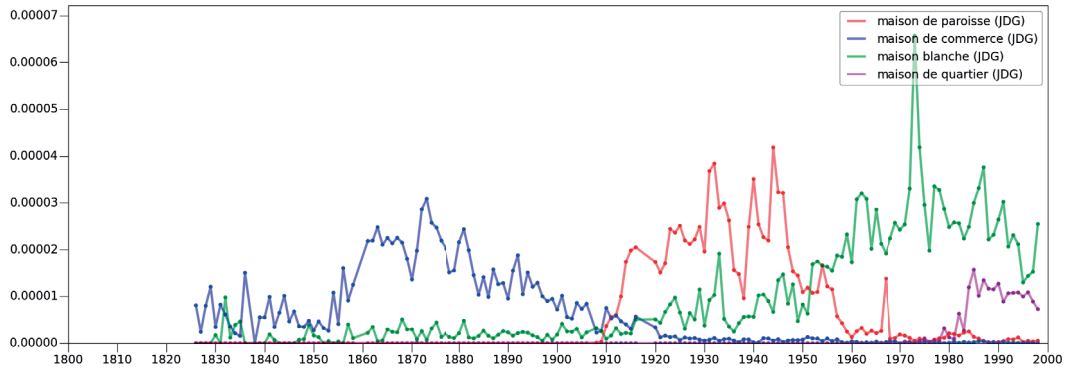


FIGURE 14.7 – Profils fréquentiels de "maison de paroisse", "maison de commerce", "maison blanche" et "maison de quartier" pour JDG

Cette disparité temporelle nous permet d'émettre l'hypothèse que la composition d'un profil fréquentiel d'un mot en une somme de profils fréquentiels de n-grammes de niveau supérieur change avec le temps et que l'étude de celle-ci permet d'éclairer de nombreux aspects sur l'évolution du mot décomposé. D'autres exemples de profils fréquentiels de mots comme "centre", "conseil", "ministre" peuvent montrer des évolutions locales intéressantes.

Quatre profils fréquentiels de n-grammes commençant par "centre" sont illustrés dans la Figure 14.8 et quatre autres commençant par "conseil" sont illustrés dans la Figure 14.9

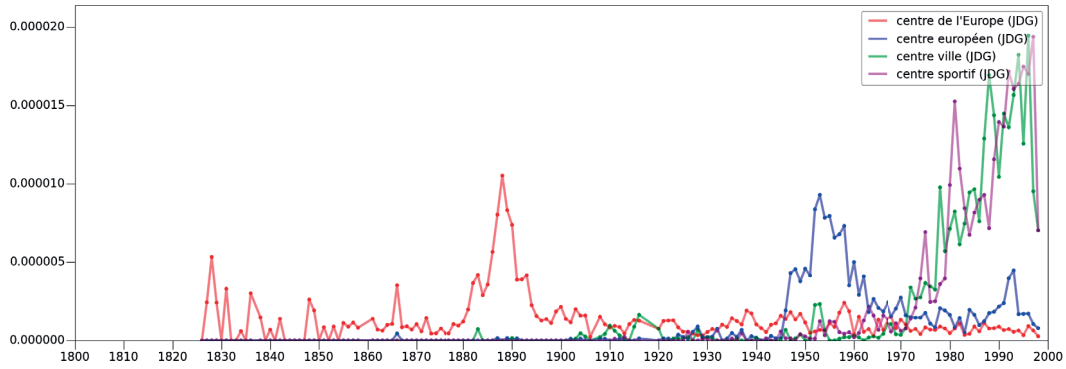


FIGURE 14.8 – Profils fréquentiels de "centre de l'Europe", "centre européen", "centre ville" et "centre sportif" pour JDG

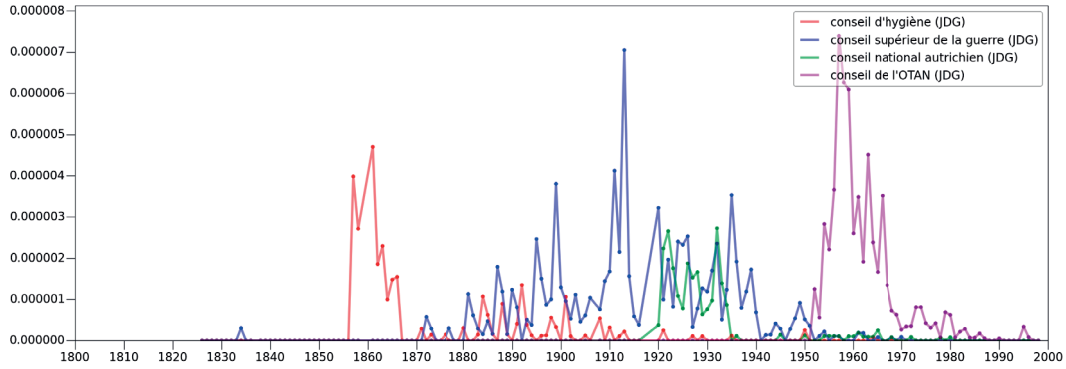


FIGURE 14.9 – Profils fréquentiels de "conseil d'hygiène", "conseil supérieur de la guerre", "conseil national autrichien", "conseil de l'OTAN" pour JDG

Des observations similaires sont constatées sur ces exemples et soulèvent la question de savoir s'il est possible de retracer l'histoire de l'utilisation d'un mot au travers de la décomposition de son profil fréquentiel selon ceux des n-grammes de niveau n plus élevé. Une façon naïve et basique de procéder à cette décomposition est de choisir un niveau n particulier et d'utiliser l'équation de décomposition sur ce niveau n. Il est ensuite possible de trier les profils fréquentiels de la décomposition en utilisant un modèle de type barycentre selon l'axe de l'année et de visualiser les profils fréquentiels ordonnés de façon à retracer de l'histoire du mot.

En continuant à prendre pour exemple la décomposition du mot "maison", nous présentons sa décomposition en 2-grammes, 3-grammes, 4-grammes et 5-grammes respectivement dans les Figures 14.10, 14.11, 14.11 et 14.12.

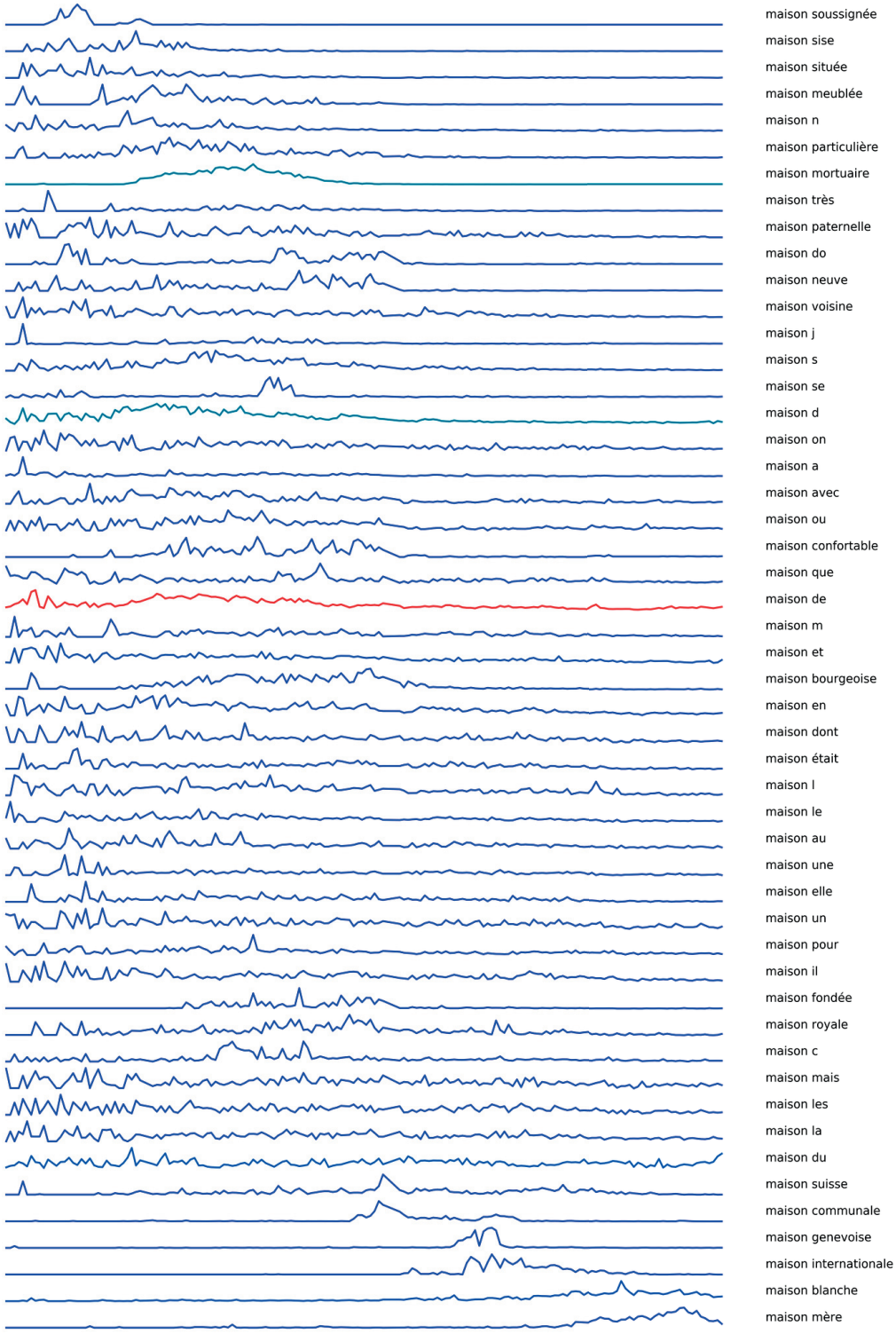


FIGURE 14.10 – Décomposition du profil fréquentiel du mot "maison" en profils fréquentiels des cinquante 2-grammes les plus fréquents, ordonnés par la moyenne pondérée des années par les fréquences de chaque profil fréquentiel (rouge : haute fréquence ; bleu : basse fréquence)

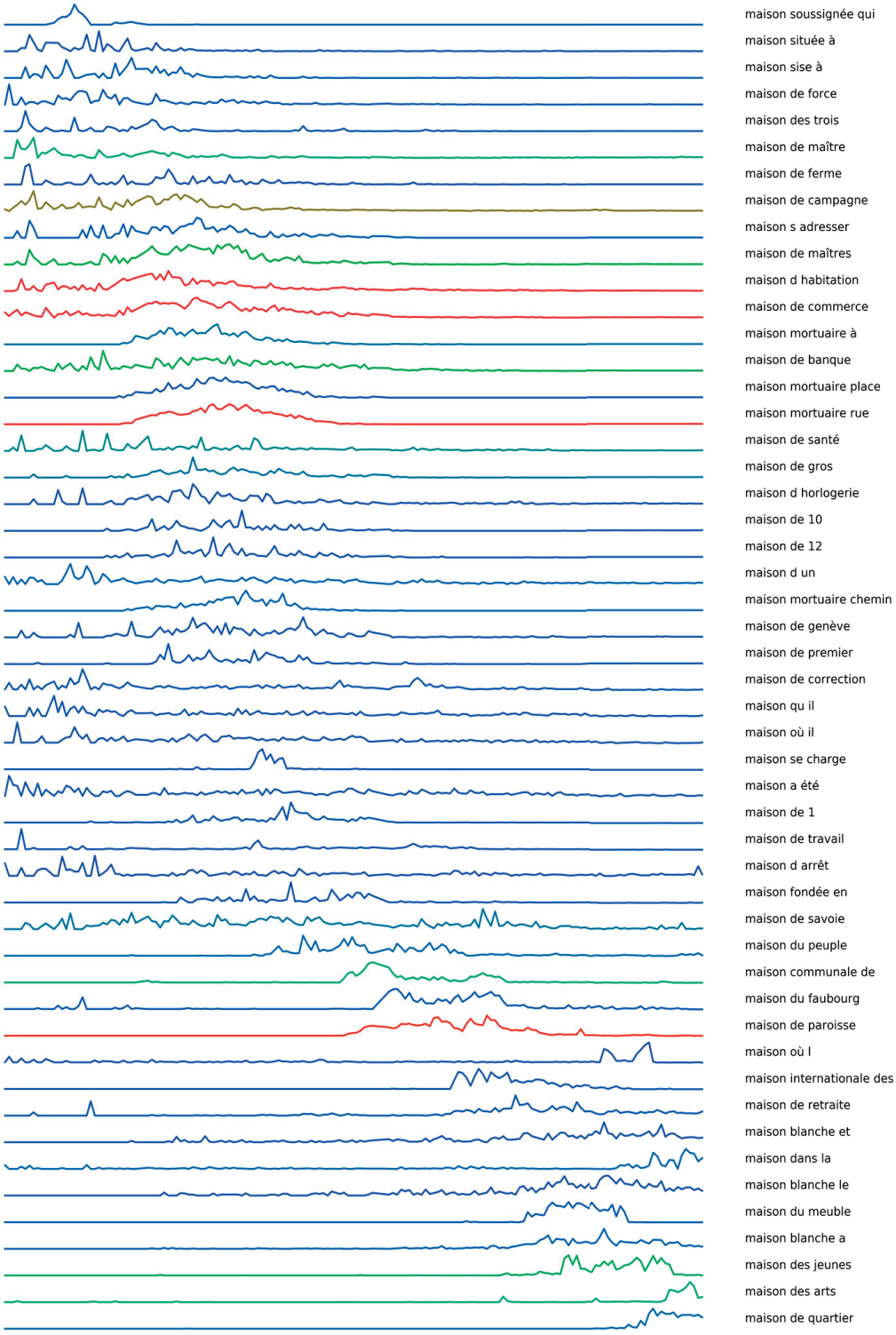


FIGURE 14.11 – Décomposition du profil fréquentiel du mot "maison" en profils fréquentiels des cinquante 3-grammes les plus fréquents, ordonnés par la moyenne pondérée des années par les fréquences de chaque profil fréquentiel (rouge : haute fréquence ; bleu : basse fréquence)

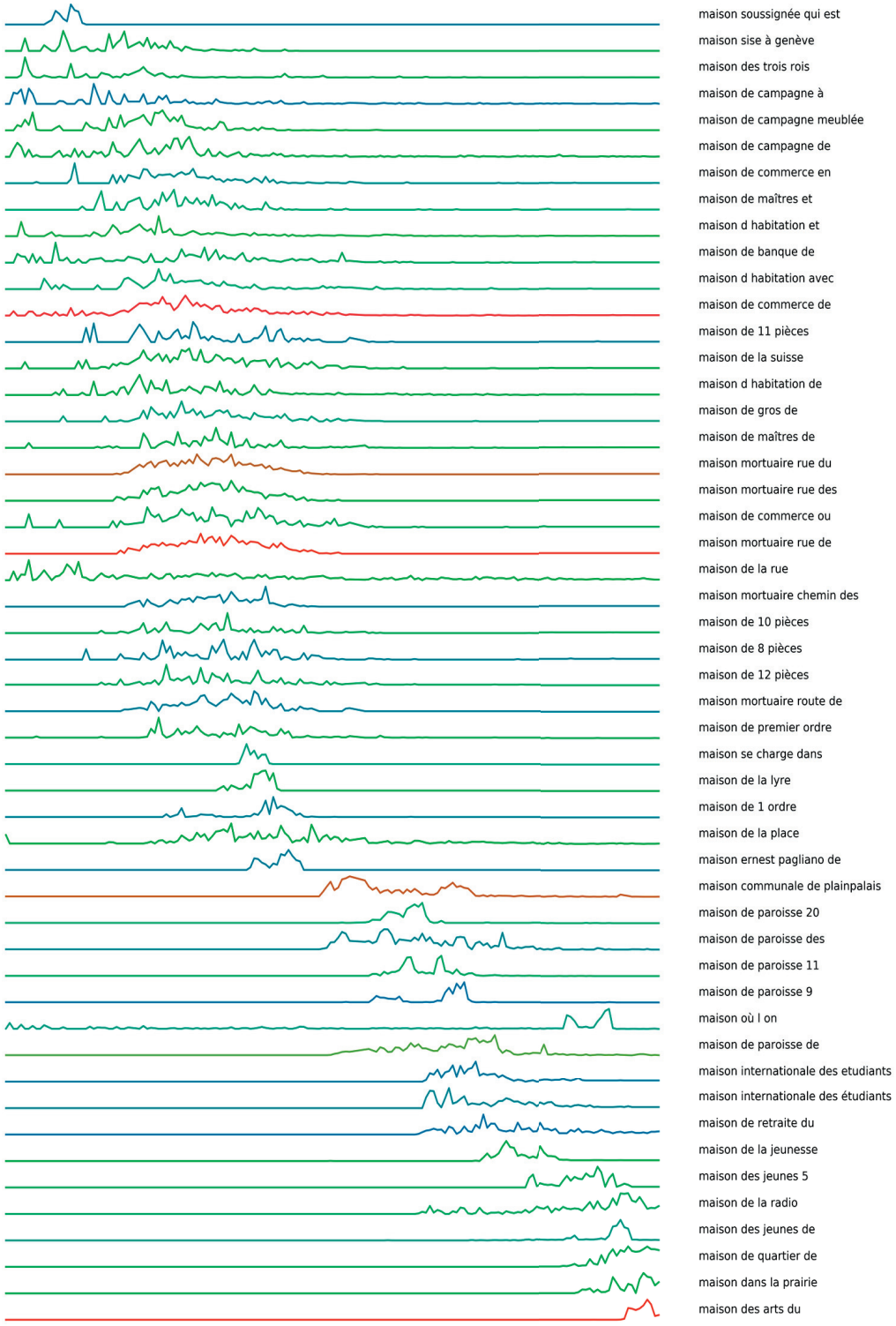


FIGURE 14.12 – Décomposition du profil fréquentiel du mot "maison" en profils fréquentiels des cinquante 4-grammes les plus fréquents, ordonnés par la moyenne pondérée des années par les fréquences de chaque profil fréquentiel (rouge : haute fréquence ; bleu : basse fréquence)

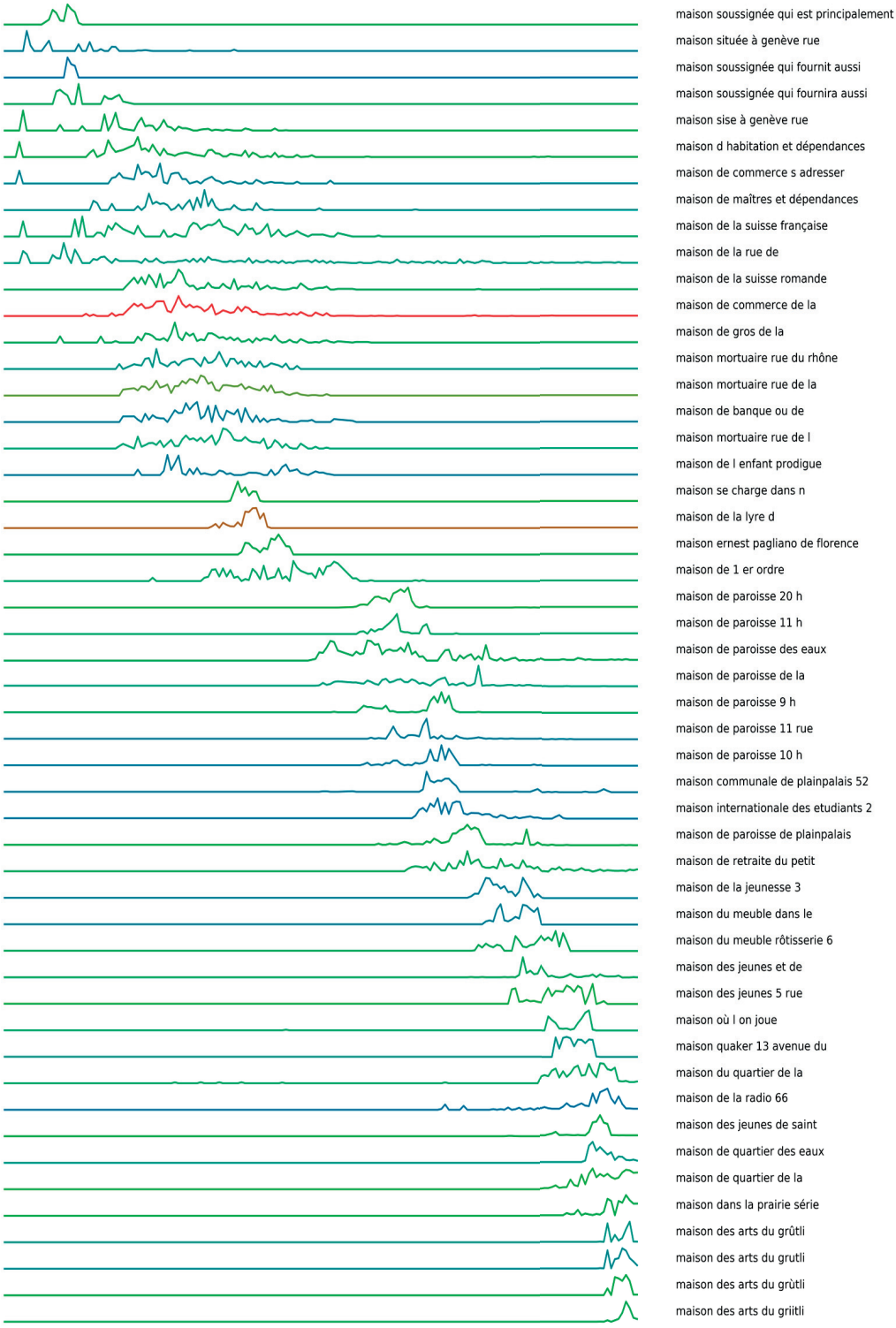


FIGURE 14.13 – Décomposition du profil fréquentiel du mot "maison" en profils fréquentiels des cinquante 5-grammes les plus fréquents, ordonnés par la moyenne pondérée des années par les fréquences de chaque profil fréquentiel (rouge : haute fréquence ; bleu : basse fréquence)

Il est passablement compliqué de repérer l'ordre temporel des profils fréquentiels dans ce type de décomposition naïve et la mesure de type barycentre que nous avons proposée est d'autant moins intuitive que le profil fréquentiel du mot analysé est dispersé dans le temps. Toutefois, nous remarquons aussi que plus la décomposition est effectuée selon des profils fréquentiels de n-grammes de niveau n élevé et plus la mesure d'ordonnancement fonctionne, car les n-grammes deviennent de plus en plus localisés dans le temps.

De façon générale et au travers des différents outils mis en place dans cette thèse, nous observons qu'un profil fréquentiel d'un n-gramme de niveau supérieur aura plus facilement une forme de courbe en cloche de type fonction gaussienne. Pour démontrer cette hypothèse nous avons ajusté des fonctions gaussiennes aux profils fréquentiels en nous basant sur 5 446 mots du noyau résilient de JDG ainsi que tous les n-grammes dont la fréquence moyenne est supérieur à 1/100 000 et dont le premier élément fait partie de l'ensemble des mots du noyau.

1 302 387 profils fréquentiels de n-grammes ont donc été ajustés sur une fonction gaussienne. Ensuite, nous appliquons un critère simple de qualité mesuré par la distance cosinus entre la courbe ajustée et les données d'origine. Basé sur ce critère, nous avons fixé un seuil de décision binaire sur la mesure cosinus à 0.8 afin que la courbe soit considérée comme bien adapté à un modèle gaussien. Les résultats de cet analyse sont présentés dans la Table 14.9.

n	Nb gaussiennes (A)	Nb de n-grammes (B)	A / B
1	311	5446	6%
2	37228	415409	9%
3	59970	518893	12%
4	48227	254277	19%
5	34991	108362	32%

TABLE 14.9 – Nombre et pourcentage de profils fréquentiels validés comme gaussien à un seuil de 0.8 sur les n-grammes de JDG dont le premier élément fait partie du noyau résilient

Le pourcentage de n-grammes correctement décrits par un modèle gaussien augmente avec le niveau n, soutenant l'hypothèse selon laquelle le modèle gaussien décrivant une courbe en forme de cloche est bien adapté à la modélisation de n-grammes de niveau supérieur. Cela signifie que lorsqu'un n-gramme est décomposé selon un niveau n supérieur, les profils fréquentiels des n-grammes de la décomposition tendent à devenir gaussiens. Cela peut être interprété comme une forme de solidification des n-grammes de niveau n supérieur, ayant leur propre trajectoire temporelle à l'intérieur de l'agrégat complexe correspondant à l'utilisation multiple d'un mot ou d'une expression donnée. Sur la base de cette hypothèse, nous utilisons le modèle gaussien afin de sélectionner les n-grammes qui seront affichés dans la décomposition et de visualiser l'histoire du mot. Nous présentons la décomposition gaussienne partielle du mot "maison" selon les profils fréquentiels des 2-grammes, 3-grammes, 4-grammes et 5-grammes gaussiens dans les Figures 14.14, 14.15, 14.16 et 14.17.

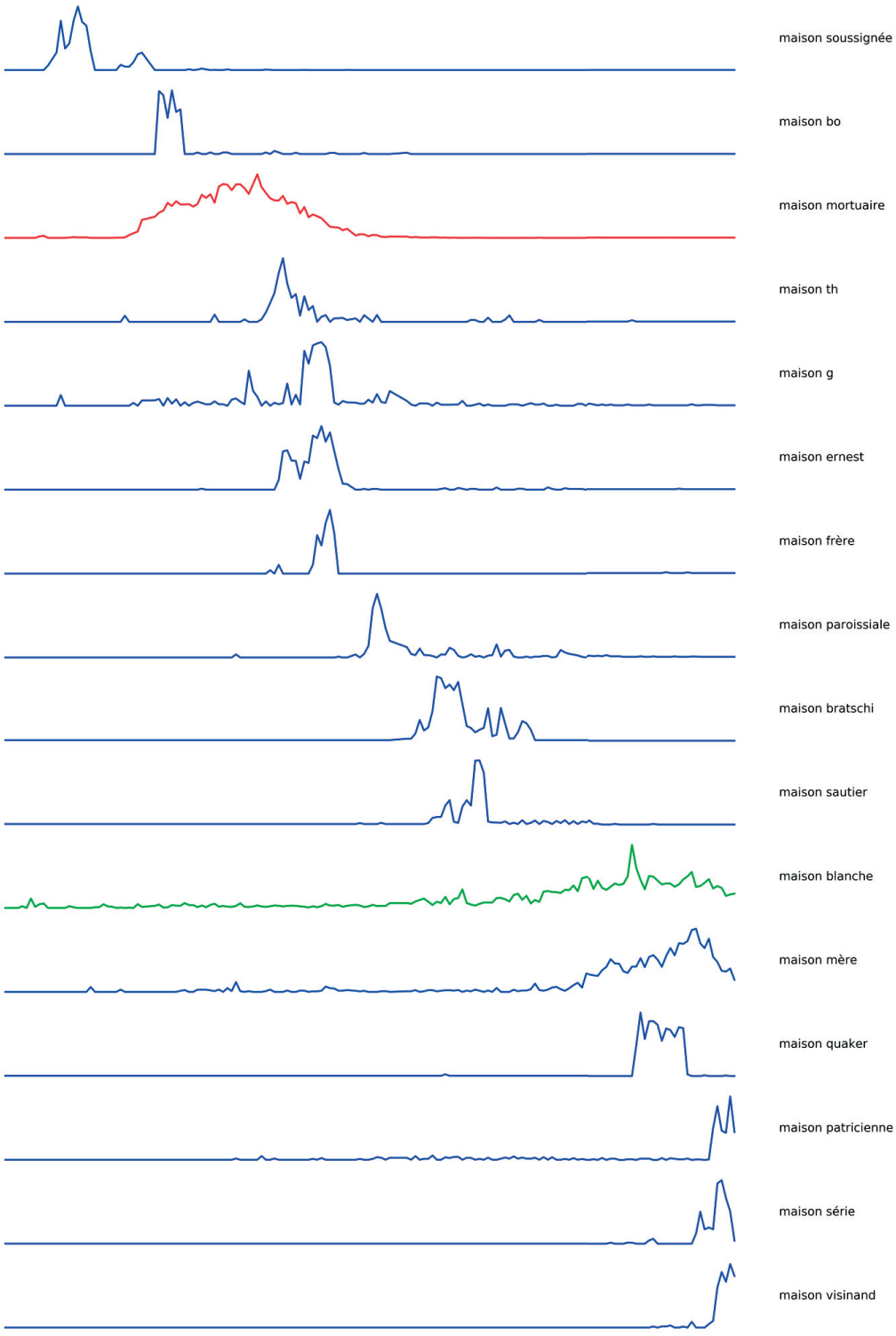


FIGURE 14.14 – Décomposition du profil fréquentiel du mot "maison" en profils fréquentiels gaussiens des 2-grammes, ordonnés par la moyenne pondérée des années par les fréquences de chaque profil fréquentiel (rouge : haute fréquence ; bleu : basse fréquence)

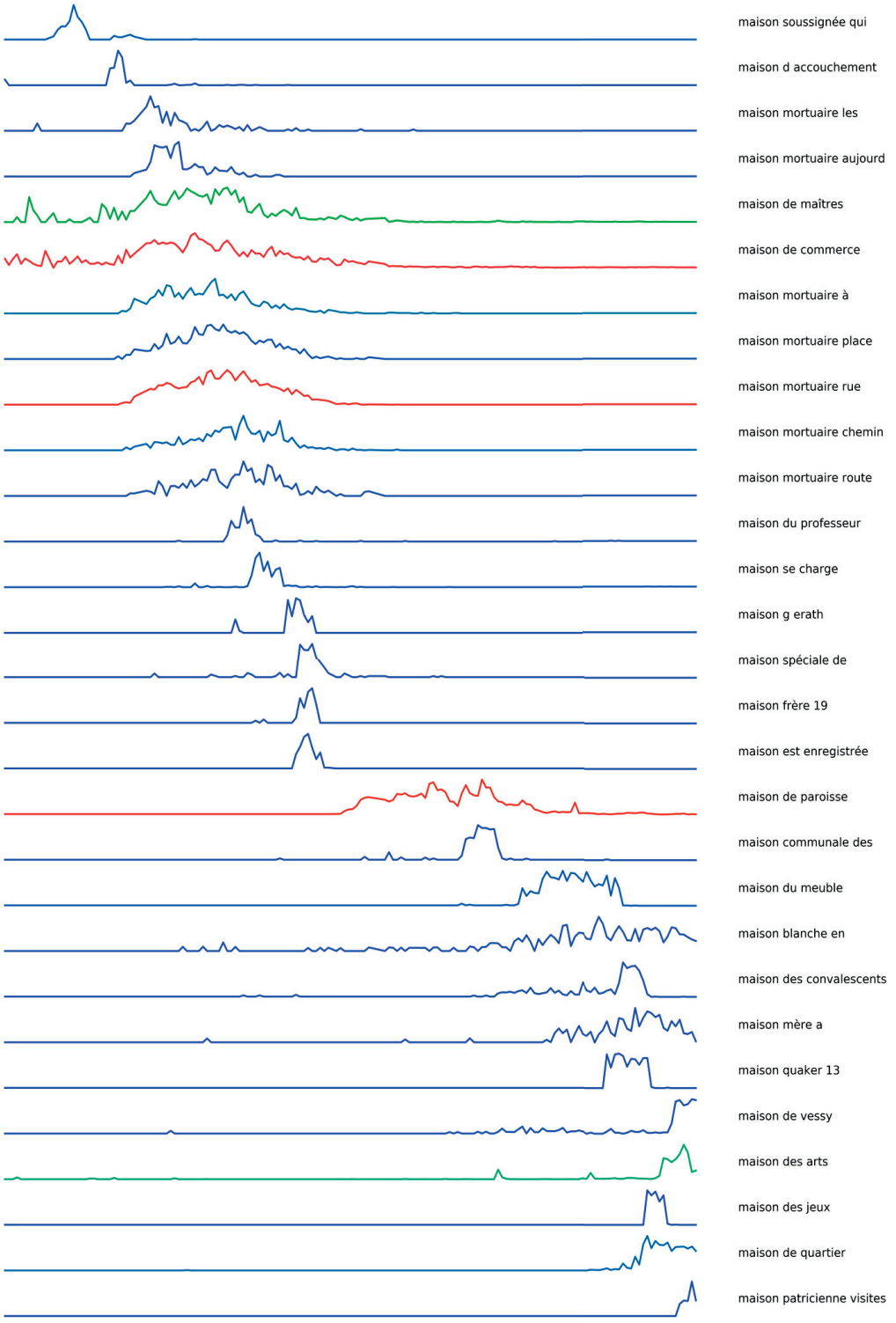


FIGURE 14.15 – Décomposition du profil fréquentiel du mot "maison" en profils fréquentiels gaussiens des 3-grammes, ordonnés par la moyenne pondérée des années par les fréquences de chaque profil fréquentiel (rouge : haute fréquence ; bleu : basse fréquence)

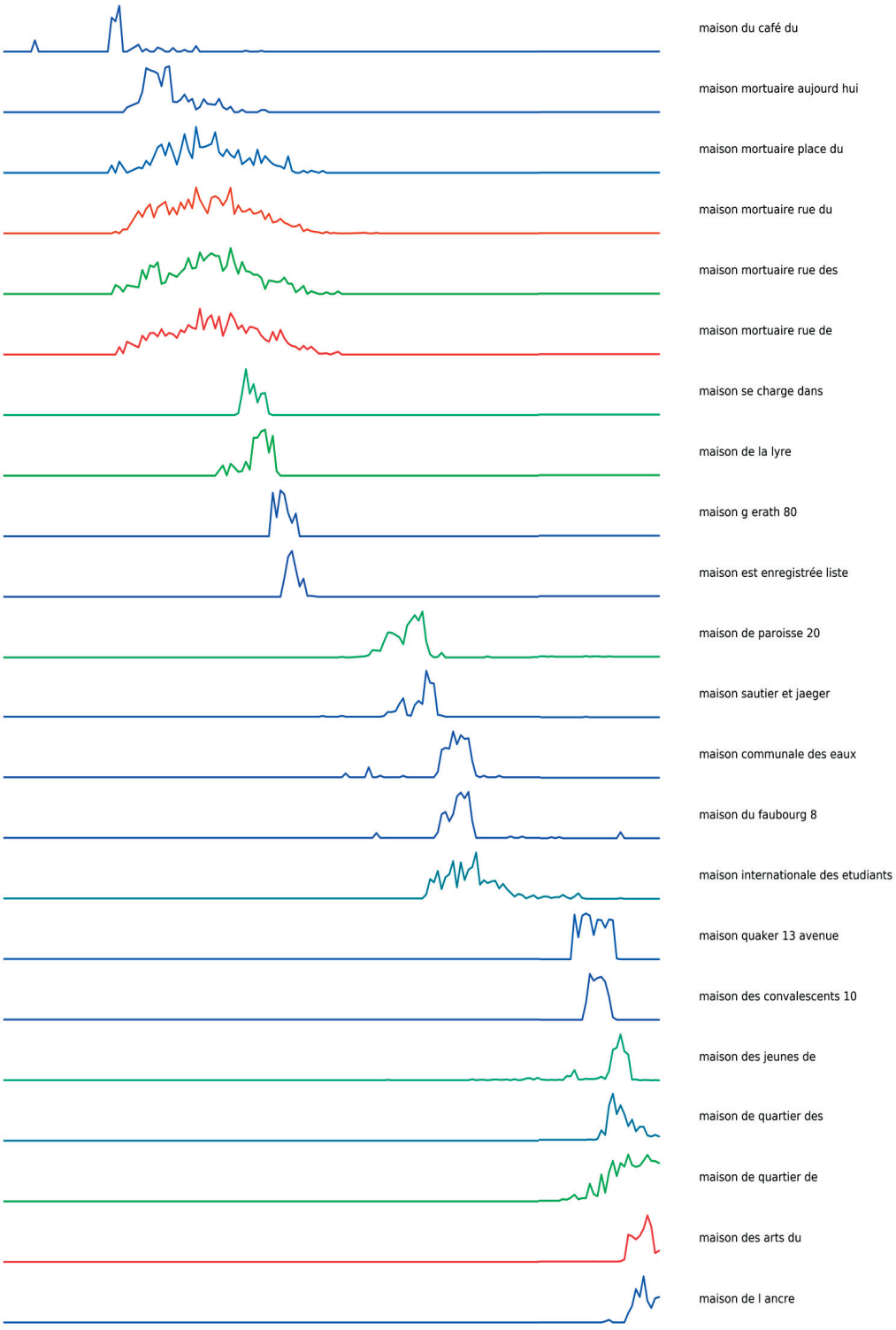


FIGURE 14.16 – Décomposition du profil fréquentiel du mot "maison" en profils fréquentiels gaussiens des 4-grammes, ordonnés par la moyenne pondérée des années par les fréquences de chaque profil fréquentiel (rouge : haute fréquence ; bleu : basse fréquence)

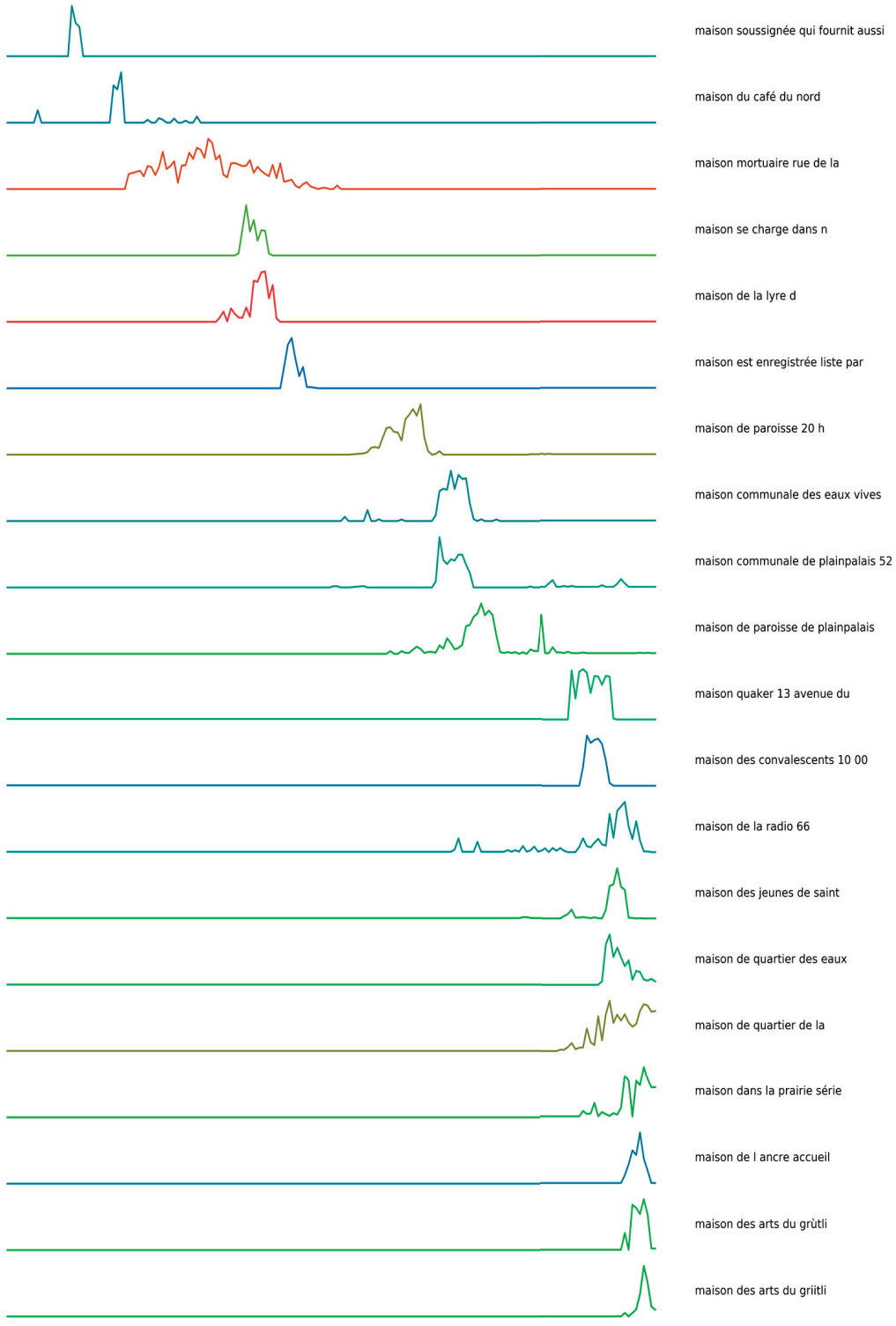


FIGURE 14.17 – Décomposition du profil fréquentiel du mot "maison" en profils fréquentiels gaussiens des 5-grammes, ordonnés par la moyenne pondérée des années par les fréquences de chaque profil fréquentiel (rouge : haute fréquence ; bleu : basse fréquence)

Ce type de décomposition permet de visualiser la temporalité locale des n-grammes et l'histoire du mot décomposé. Toutefois, il faut remarquer que cette décomposition n'est pas complète puisque seuls les n-grammes dont le profil fréquentiel est considéré gaussien sont affichés. Cependant, il est possible de jouer sur le seuil de similarité entre le profil fréquentiel et la fonction gausienne qui lui est ajustée. Une réduction du seuil permet d'accepter des profils plus approximativement gaussiens tandis qu'une augmentation aura l'effet inverse.

Chaque niveau de décomposition est composée de n-grammes intéressants et des expressions figées correspondant aux expressions multi-mots font leur apparition si ce n'est que l'on constate aussi qu'une expression comme "maison de quartier de" ne doit sa forme gaussienne qu'au fait que l'expression figée "maison de quartier" en est elle même une. Ainsi, beaucoup d'expressions peu intéressantes pour la décomposition sont donc également présentes, car la décomposition est faite "à l'aveugle" acceptant toutes formes gausiennes d'un niveau particulier n.

Nous constatons, par exemple dans le cas du 3-gramme "maison de paroisse", que si l'on continue à monter le niveau n, on peut obtenir un sens encore plus précis puisque le mot "de" est utilisé afin de donner une précision sémantique supplémentaire, mais uniquement à la connaissance de la suite des mots qui le suivent. Ainsi le 4-gramme "maison de paroisse de" ne donne pas de précision supplémentaire, mais prépare le terrain pour les niveaux n plus élevés et le 5-gramme "maison de paroisse de Plainpalais" se réfère directement à une maison de paroisse particulière de Genève située dans le quartier de Plainpalais.

Dans le cas particulier du mot "maison", plusieurs n-grammes de niveau n élevé finissent par donner une adresse. C'est notamment dû au fait de diverses annonces permettant à ces n-grammes de niveaux supérieurs d'être plus fréquents. Toutefois, cet effet reste marginal dans la plupart des cas. Par exemple, dans le cas des n-grammes "maison de paroisse de Plainpalais" et "maison de paroisse", le profil fréquentiel du premier forme une gaussienne d'amplitude plus faible, de largeur plus faible et dont le centre est différent de celui du deuxième. Ces exemples nous permettent d'observer de manière directe comment un n-gramme a tendance à devenir de plus en plus localisé dans le temps en progressant dans les niveaux n et notamment à quel point le phénomène est rapide (cf. Figure 12.1), dès $n = 3, 4$ et 5.

Afin d'obtenir une décomposition plus complète et composée de n-grammes localisés dans le temps, nous utilisons le principe de la décomposition minimale (cf. partie théorique) nous permettant de retrouver une cohérence dans la décomposition en excluant les profils fréquentiels gaussiens de n-grammes commençant par un (n-1)-gramme déjà gaussien.

De plus, la décomposition minimale permet de transcender le niveau n des n-grammes afin d'obtenir une décomposition multi-échelle. Le modèle gaussien est bien adapté pour décrire le phénomène des n-grammes en décomposition, mais il est possible d'intégrer également d'autres modèles. Le principe de la décomposition minimale reste le même quel que soit le modèle et celui ci joue donc le rôle d'un catalyseur permettant de choisir des n-grammes au profil fréquentiel intéressant comme faisant partie de la décomposition multi-échelle.

La modélisation des profils fréquentiels par des gausiennes a été stockée dans une base de données nous permettant de créer un outil online afin de produire la décomposition en temps réel sur la base d'un mot requis par l'utilisateur. Une illustration d'une partie des résultats de la décomposition minimale de "maison" et "centre" obtenus à l'aide de l'outil online est présenté dans la Figure 14.18. Les profils fréquentiels des n-grammes sont accompagnés de leur fonction gaussienne et le premier graphe représente le profil fréquentiel du n-gramme décomposé ainsi que la contribution des n-grammes de la décomposition.

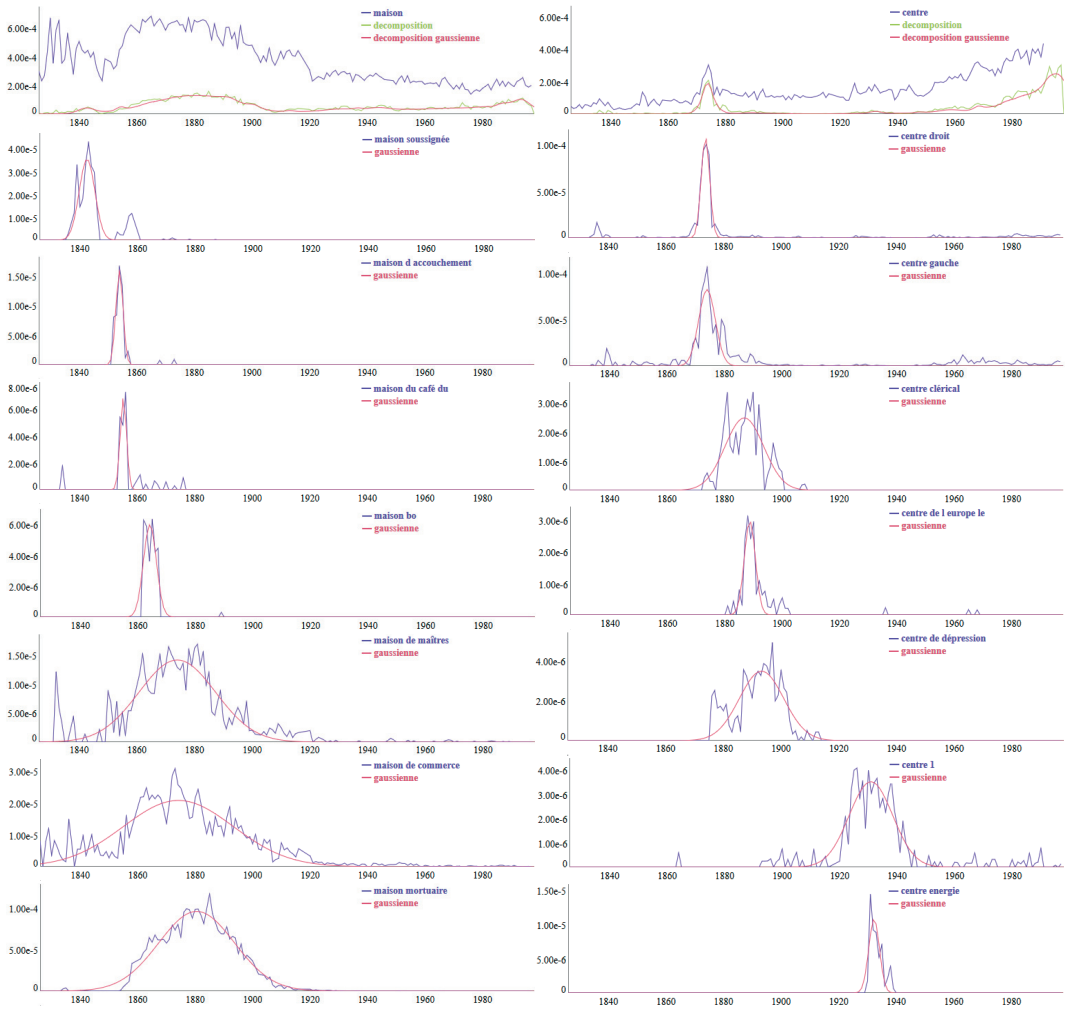


FIGURE 14.18 – Outil online de décomposition minimale d'un mot requis par l'utilisateur avec possibilité de changer le seuil de similarité requis du modèle gaussien

Nous présentons, sous un format différent, la décomposition minimale gaussienne multi-échelle complète des mots "maison", "centre", "conseil", "ministre" et "relativement" respectivement dans les Figures 14.19, 14.20, 14.21, 14.22, 14.23. Le seuil a été baissé de 0.8 à 0.7 pour "ministre" et de 0.8 à 0.5 pour "relativement" afin d'afficher des profils fréquentiels moins gaussiens, mais qui ont tout de même une forme proche d'une courbe en cloche.

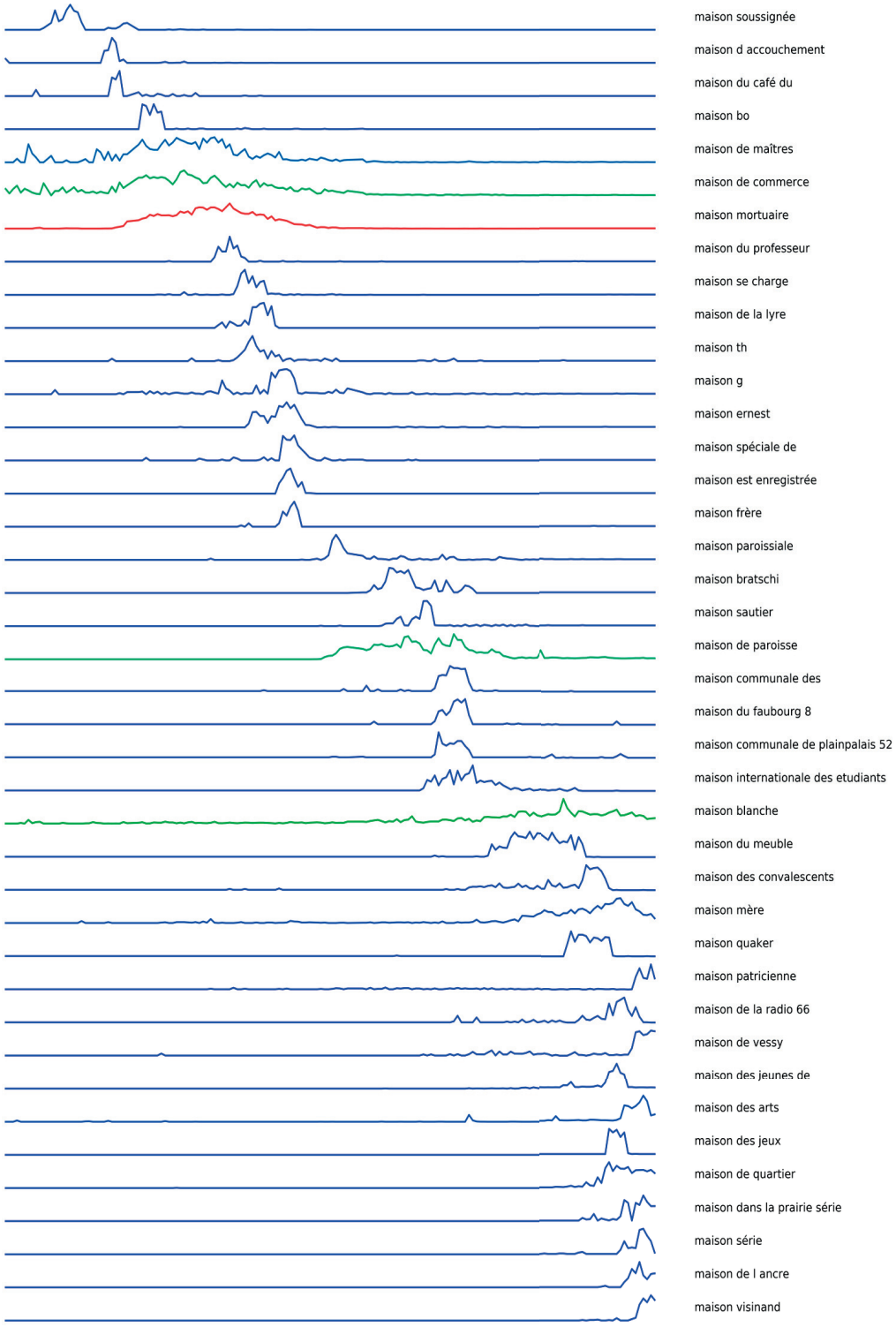


FIGURE 14.19 – Décomposition minimale du profil fréquentiel du mot "maison" en profils fréquentiels des n-grammes modélisé par la fonction gaussienne, ordonnés par la moyenne pondérée des années par les fréquences de chaque profil fréquentiel (rouge : haute fréquence ; bleu : basse fréquence)

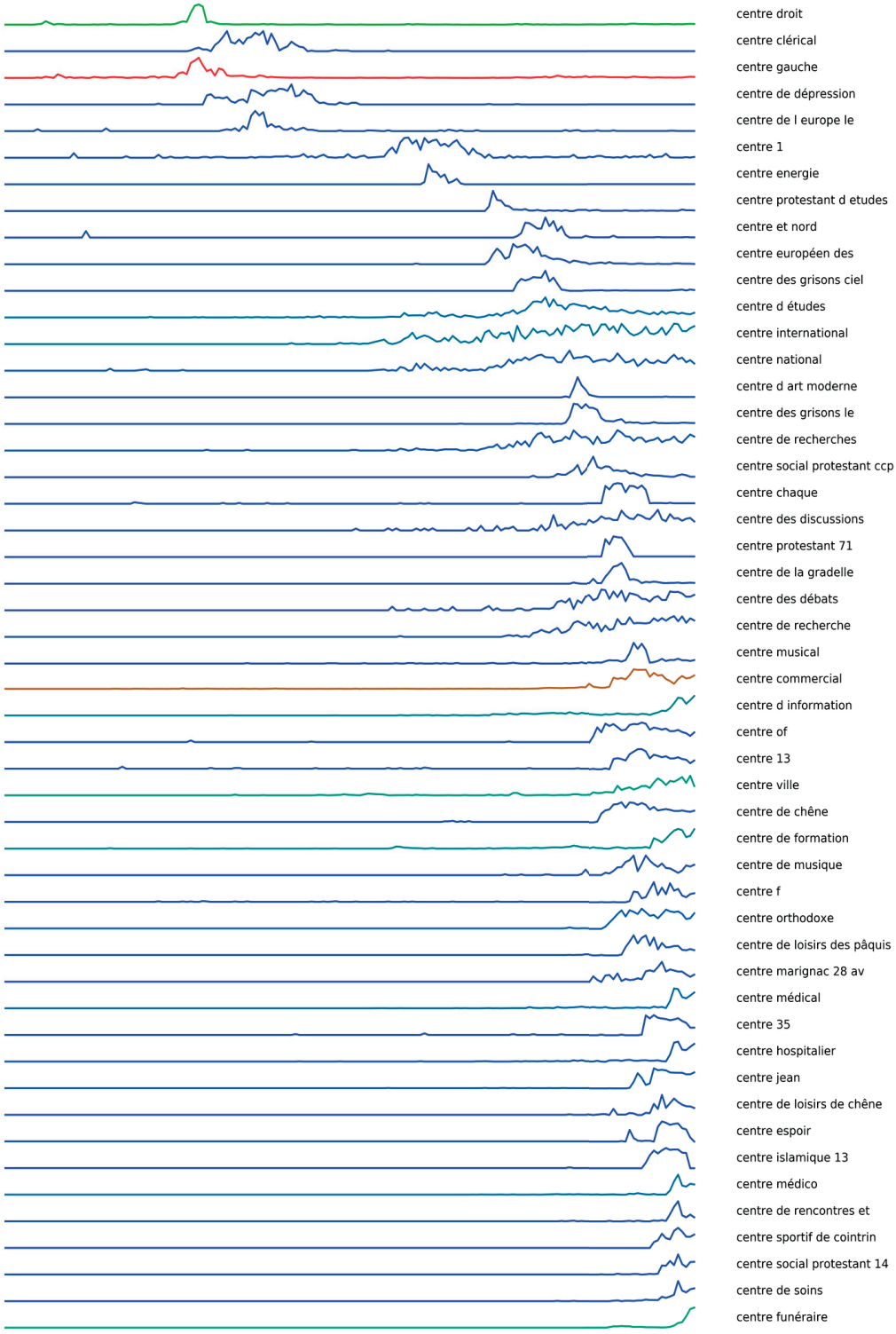


FIGURE 14.20 – Décomposition minimale du profil fréquentiel du mot "centre" en profils fréquentiels des n-grammes modélisé par la fonction gaussienne, ordonnés par la moyenne pondérée des années par les fréquences de chaque profil fréquentiel (rouge : haute fréquence ; bleu : basse fréquence)

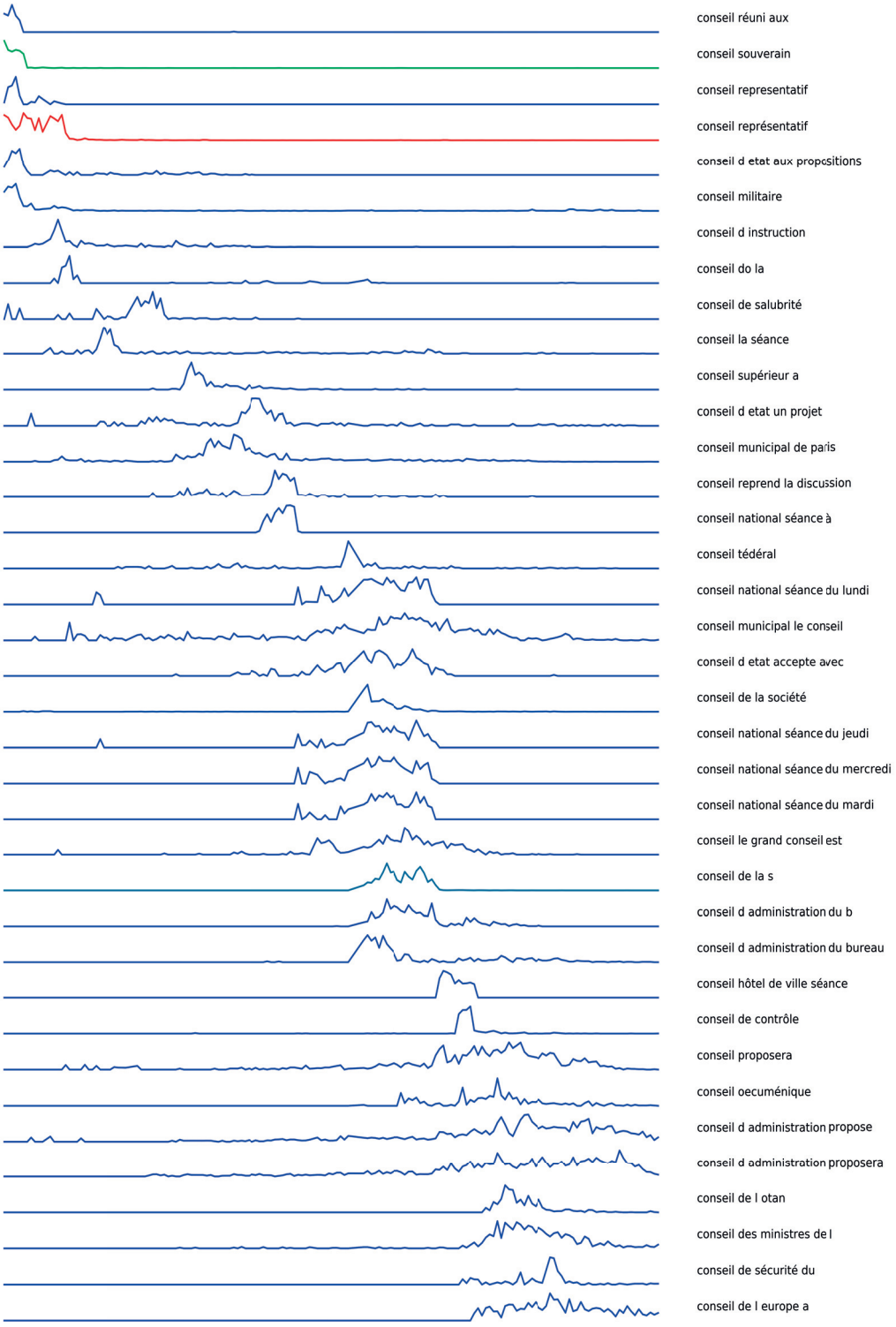


FIGURE 14.21 – Décomposition minimale du profil fréquentiel du mot "conseil" en profils fréquentiels des n-grammes modélisé par la fonction gaussienne, ordonnés par la moyenne pondérée des années par les fréquences de chaque profil fréquentiel (rouge : haute fréquence ; bleu : basse fréquence)

FIGURE 14.22 – Décomposition minimale du profil fréquentiel du mot "ministre" en profils fréquentiels des n-grammes modélisé par la fonction gaussienne, ordonnés par la moyenne pondérée des années par les fréquences de chaque profil fréquentiel (rouge : haute fréquence ; bleu : basse fréquence)

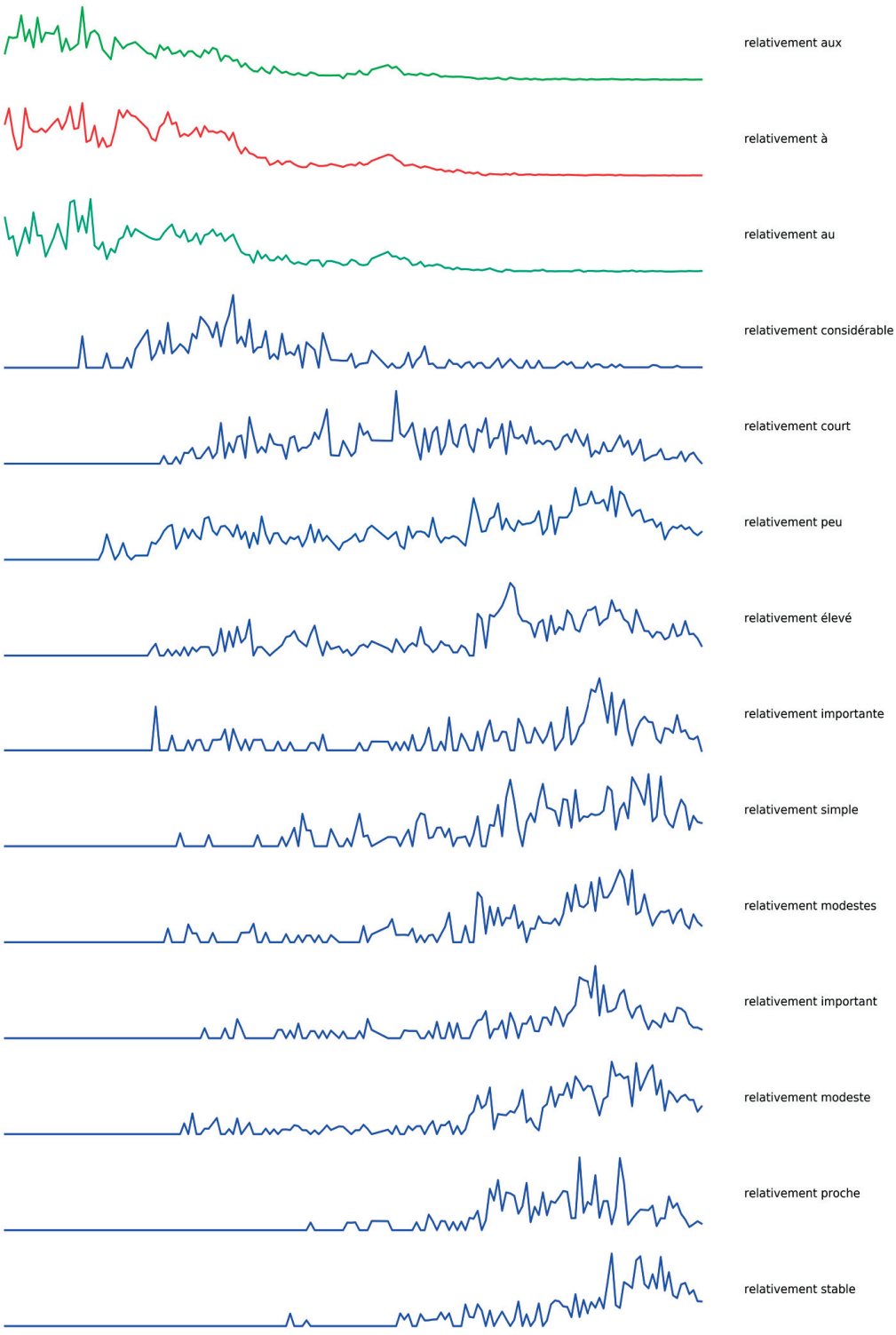


FIGURE 14.23 – Décomposition minimale du profil fréquentiel du mot "relativement" en profils fréquentiels des n-grammes modélisé par la fonction gaussienne, ordonnés par la moyenne pondérée des années par les fréquences de chaque profil fréquentiel (rouge : haute fréquence ; bleu : basse fréquence)

Nous observons sur la décomposition du mot "maison" que les n-grammes de profils fréquentiels de type gaussien les plus fréquents sont "maison de commerce", "maison mortuaire", "maison de paroisse" et "maison blanche". Ainsi "maison de commerce" et "maison mortuaire" furent principalement utilisés dans le journal de Genève entre 1855 et 1910. "maison de paroisse" fut utilisé dès 1910 jusque 1970 tandis que "maison blanche" est apparu de façon progressive, mais décolle dès 1915 et atteint un régime de stabilité dans les dernières années du corpus. Les profils fréquentiels des n-grammes "maison de commerce", "maison mortuaire", "maison de paroisse" et "maison blanche" sont présentés dans la Figure 14.24.

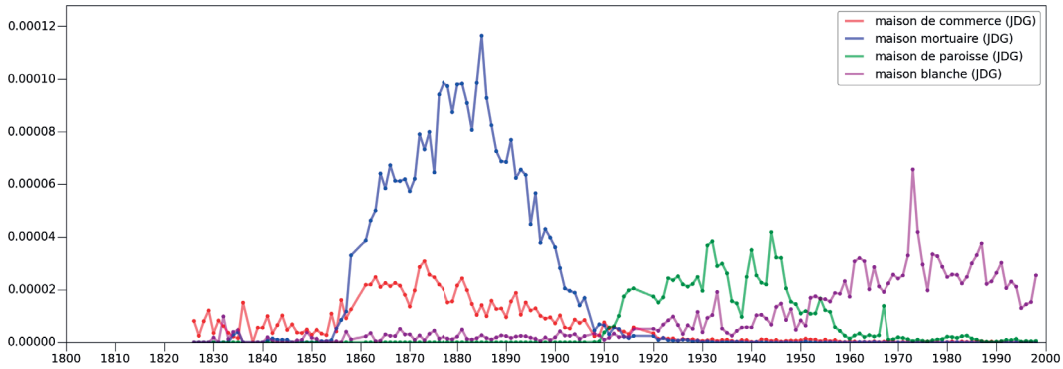


FIGURE 14.24 – Profils fréquentiels de "maison de commerce", "maison mortuaire", "maison de paroisse" et "maison blanche" pour JDG

La décomposition du mot "centre" met en évidence deux 2-grammes de fréquence élevée, "centre droit" et "centre gauche". Etonnamment ces termes sont utilisés massivement de 1869 à 1893 pour décrire une orientation politique selon la traditionnelle projection sur une droite de dimension 1. Leurs profils fréquentiels sont spécifiquement localisés dans le temps et l'utilisation de ces termes s'est ensuite rapidement détériorée, le phénomène étant similaire dans les deux journaux. Les profils fréquentiels des n-grammes "centre gauche" et "centre droit" sont présentés dans la Figure 14.25 pour les corpus de JDG et GDL.

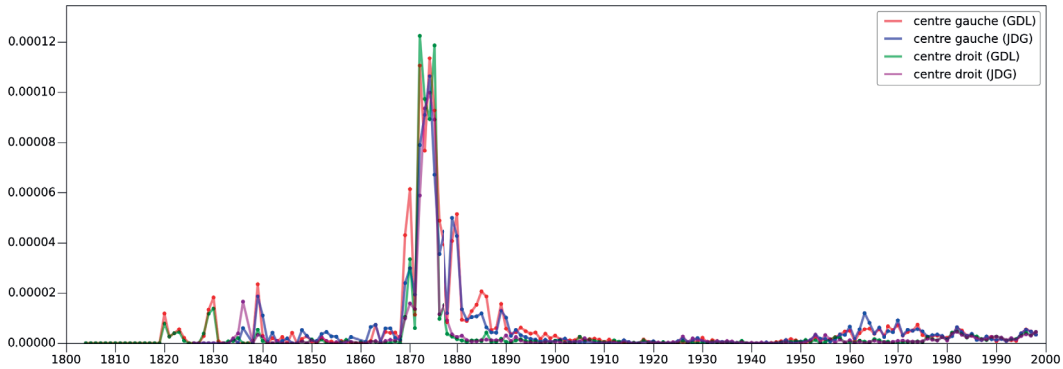


FIGURE 14.25 – Profils fréquentiels de "centre gauche" et "centre droit"

Les autres n-grammes de fréquence élevée présents dans la décomposition de "centre" sont "centre commercial", "centre ville", "centre de formation" et "centre funéraire" dont les utilisations sont particulièrement récentes. Les profils fréquentiels des n-grammes "centre commercial", "centre ville", "centre de formation" et "centre funéraire" sont présentés dans la Figure 14.26.

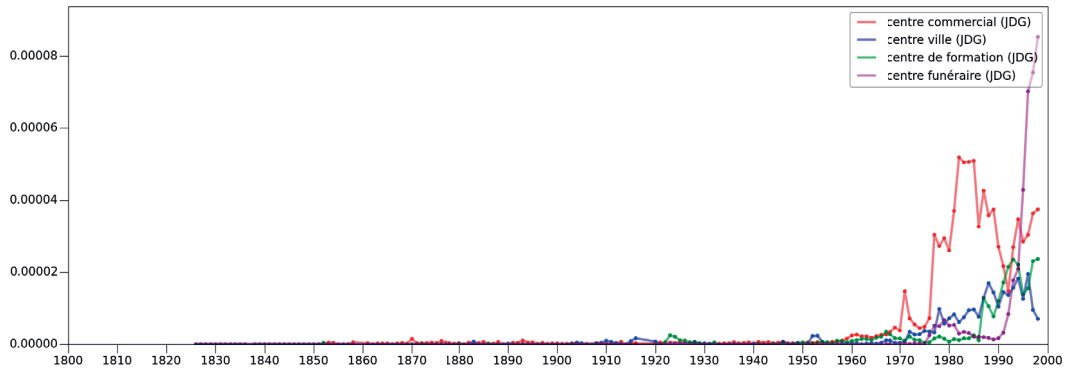


FIGURE 14.26 – Profils fréquentiels de "centre commercial", "centre ville", "centre de formation" et "centre funéraire" pour JDG

Dans le cas de la décomposition du mot "conseil", deux 2-grammes ont une fréquence élevée par rapport aux autres, il s'agit de "conseil souverain" et "conseil représentatif". Ces 2-grammes furent utilisés dans les années les plus anciennes du corpus et principalement par le journal de JDG. Les profils fréquentiels des n-grammes "conseil souverain" et "conseil représentatif" sont présentés dans la Figure 14.27.

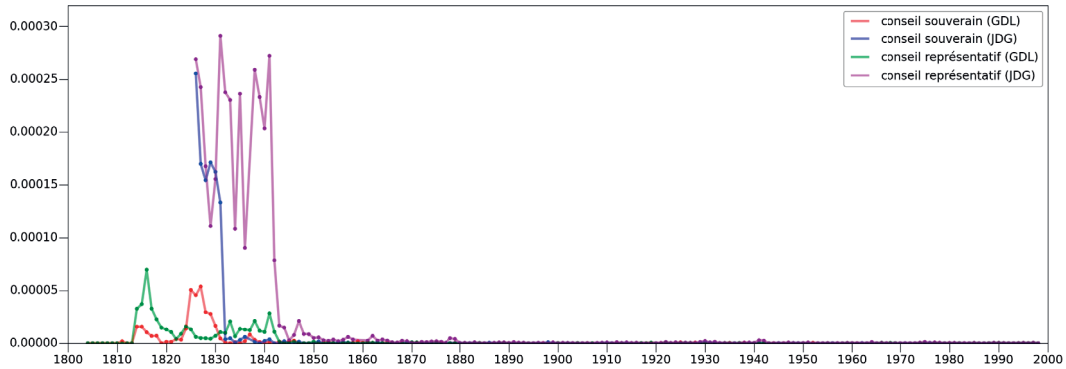


FIGURE 14.27 – Profils fréquentiels de "conseil souverain" et "conseil représentatif"

La décomposition du mot "ministre" met en évidence différents n-grammes permettant de préciser le ministère auquel appartient le ministre mentionné. Ainsi on observe de nombreux n-grammes commençant par les 2-grammes "ministre de" ou "ministre du".

Les n-grammes de profil fréquentiel gaussien les plus fréquents sont "ministre de Suisse", "ministre de la défense", "ministre de l'économie" et "ministre israélien". Leur profils fréquentiels sont présentés dans la Figure 14.28.

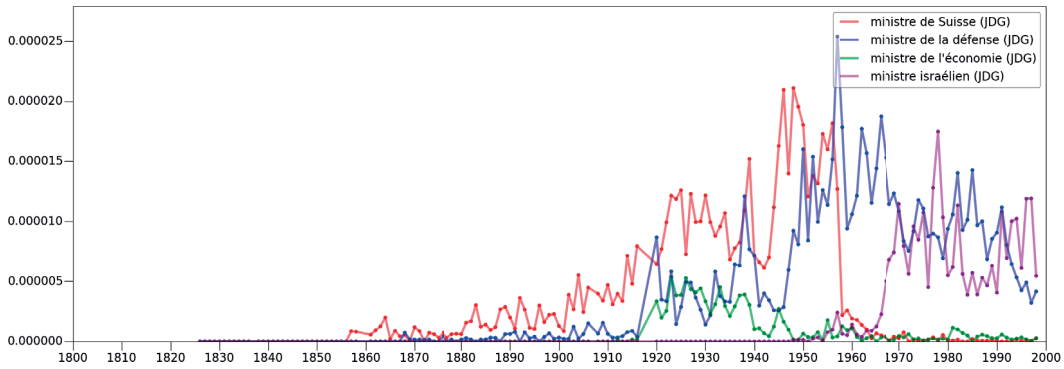


FIGURE 14.28 – Profils fréquentiels de "ministre de Suisse", "ministre de la défense", "ministre de l'économie" et "ministre israélien" pour JDG

L'exemple de la décomposition du mot "relativement" est emblématique, car il révèle de façon automatique une évolution linguistique constatée auparavant au travers d'une décomposition manuelle dans le corpus de GDL (cf. Figures 12.70 et 12.72).

Les profils fréquentiels des n-grammes "relativement", "relativement aux", "relativement à" et "relativement au" sont présentés dans la Figure 14.29.

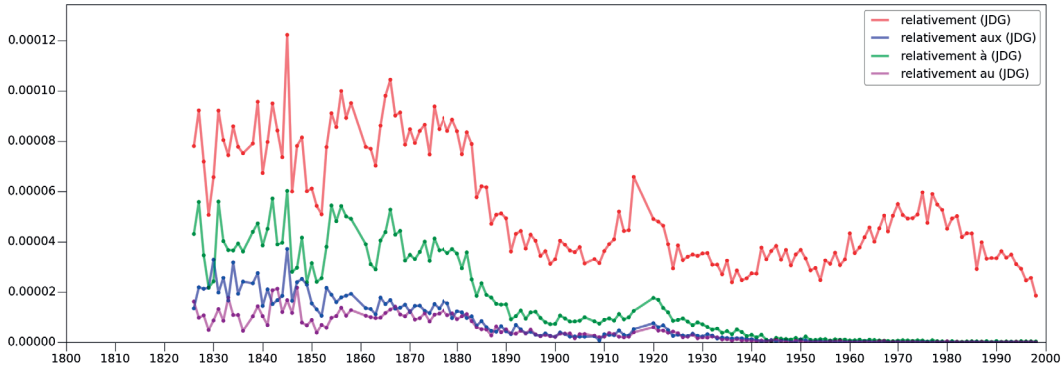


FIGURE 14.29 – Profils fréquentiels de "relativement", "relativement aux", "relativement à" et "relativement au" pour JDG

Nous observons l'évolution des trois 2-grammes qui dans les années anciennes sont les composantes principales et quasiment uniques du profil fréquentiel du mot "relativement". Nous observons toutefois que plus aucun de ces trois 2-grammes n'est utilisé dès 1940.

les profils fréquentiels des n-grammes "relativement considérable", "relativement court", "relativement peu" et "relativement élevé" sont présentés dans la Figure 14.30.

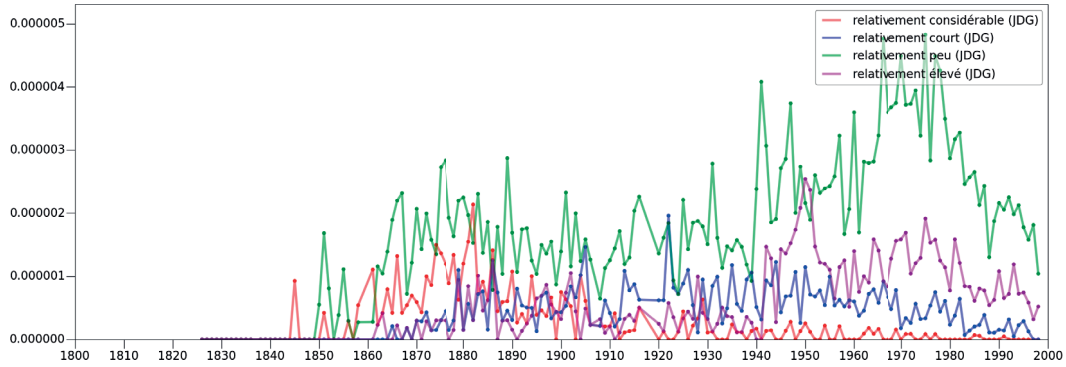


FIGURE 14.30 – Profils fréquentiels de "relativement considérable", "relativement court", "relativement peu" et "relativement élevé" pour JDG

A l'inverse, nous observons l'apparition d'une pléthore de 2-grammes de faible fréquence composés du mot "relativement" suivi d'un adjectif. La décomposition dans ce cas confirme une évolution sémantique du mot "relativement" qui signifie "à propos de", "quant à" ou "au sujet de" dans les années anciennes avant de signifier plutôt "passablement", "jusqu'à un certain point" dans les années les plus récentes.

En résumé, ces décompositions multi-échelle minimales, au travers de la modélisation gaussienne, semblent fournir une décomposition intéressante en terme d'expressions multi-mots et expressions figées, autonomes dans la langue. Elle est tout de même imparfaite, car certaines expressions ne correspondent pas à cette typologie ou se terminent par le mot "et" ou "de" traduisant le fait que l'expression n'est pas encore totalement figée ou autonome.

En effet, les n-grammes non figés comme "maison de" ont un profil fréquentiel similaire à leur ascendant "maison" et ils n'ajoutent pas beaucoup d'information sur l'utilisation du mot "maison" dans le corpus. Cependant, l'autonomie des expressions multi-mots comme le 3-gramme "maison de paroisse" ou le 5-gramme "maison de paroisse de plainpalais" s'observe par une différence élevée du comportement des profils fréquentiels par rapport à leurs ascendants respectifs "maison de" et "maison de paroisse de".

Un seuil de 0.8 doit être dépassé par la similarité entre la modélisation gausienne et le profil fréquentiel considéré afin qu'un n-gramme soit inclus dans la décomposition. Ce seuil est arbitraire et peut être changé par l'utilisateur de l'outil online afin de réduire ou augmenter le conditionnement au modéle gaussien.

Toutefois, la valeur de 0.8 a été choisie en fonction des résultats appliqués aux 1 302 387 n-grammes dont le premier élément fait partie du noyau résilient (cf. Table 14.9) et donne empiriquement des résultats intéressants pour la décomposition.

En utilisant ces décompositions, nous montrons que nous pouvons comprendre une partie de l'histoire des mots au travers d'une modélisation des profils fréquentiels et d'une décomposition multi-échelle minimale. Nous observons que le modèle des gaussiennes en forme de courbe en cloche est particulièrement efficace dans l'approche multi-échelle de la décomposition. Nous avons constaté que bon nombre de profils fréquentiels de n-grammes de type gaussien correspondent à une "solidification" ou "figement" linguistique et sont susceptibles d'être qualifiés d'expressions multi-mots ou expressions figées.

Cette méthode ouvre donc une voie permettant de détecter automatiquement les expressions multi-mots en fonction de l'évolution fréquentielle des n-grammes. Nous avons également développé un outil interactif pour visualiser en temps réel la décomposition en termes de gaussiennes minimales. Cet outil permet de jouer sur le seuil de similarité pour le modèle gaussien et permet d'étudier l'histoire des mots en terme de n-grammes de niveau n supérieurs.

Les exemples montrés sont principalement liés au journal de JDG, mais le corpus de GDL a également été modélisé sur la même base que celui de JDG. La comparaison de l'histoire des mots dans les deux journaux offre des informations supplémentaires sur les effets potentiellement linguistiques ou plutôt liées au corpus lui-même. En outre, cela permet de détecter les évolutions qui sont propres ou communes aux deux journaux.

Ce chapitre a exploré une façon particulière d'analyser la relation entre un n-gramme et les n-grammes de niveau supérieur qui contiennent celui-ci. Il se concentre sur les n-grammes partageant un ou plusieurs mots en communs. En effet, le même raisonnement est valable sur une base rétrospective plutôt que prospective. Dans la vision prospective, le texte se déploie dans l'ordre d'écriture comme nous l'avons analysé dans ce chapitre et est plus intuitif. Toutefois, il est équivalent de modéliser la décomposition des n-grammes selon la vision rétrospective de l'équation de décomposition, en considérant comme base le mot qui termine les n-grammes. Il pourrait être intéressant d'utiliser d'autres modélisations des n-grammes de niveau supérieur ainsi que de tester des décompositions minimales multi-modèles.

Dans tous les cas, la méthode de décomposition minimale produit des résultats automatiques susceptibles de mettre rapidement en évidence un changement linguistique sur de grands corpus et est indépendante du langage au même titre que les autres outils développés dans cette thèse. Une production automatique de ces résultats pour l'ensemble de tous les mots du corpus et l'outil en ligne permettraient alors à un linguiste ou un chercheur en linguistique d'analyser les évolutions linguistiques du corpus avec une plus grande efficacité et précision.

Dans la section suivante, nous tentons d'étudier les relations de n-grammes qui ne sont pas liés par l'équation de décomposition. Il s'agit donc de l'étude de l'espace complémentaire.

14.4 Espace des corrélations fréquentielles

Dans cette section, nous utilisons la notion mathématique simple de corrélation afin de reconstruire un nouvel espace métrique de n-grammes sur la base des profils fréquentiels individuels de ceux-ci. Bien que cette méthode puisse être appliquée à chaque niveau individuel n, elle est également généralisable à un niveau multi-échelle.

Afin d'éliminer les relations particulières de n-grammes liés par l'équation de décomposition des profils fréquentiels, nous considérons uniquement la relation entre deux n-grammes qui n'ont aucune chaîne de mots en commun.

Les profils fréquentiels sont des séries chronologiques, mais aussi au sens plus général des courbes et, mathématiquement, des fonctions envoyant une abscisse entière x représentant une année sur une ordonnée réelle y représentant une fréquence relative. Une mesure basique de la relation entre deux courbes n'accordant pas d'importance au niveau fréquentiel moyen, est donnée par le calcul du coefficient de corrélation de Pearson (Galton, 1886).

De façon computationnelle, la méthode est simple à mettre en oeuvre, mais elle requiert de sauvegarder un grand volume de données vu qu'une comparaison deux à deux augmente la quantité de données selon le carré du nombre de courbes comparées. Nous sélectionnons donc l'espace particulier des mots résilients de plus de 150 années auquel nous ajoutons l'espace des n-grammes sous-tendus par ces mots. L'espace est donc constitué des n-grammes dont le premier mot est inclus dans l'ensemble de résilience 150 du niveau $n = 1$.

Nous avons créé un outil de recherche online permettant d'extraire en temps réel les n-grammes les plus corrélés à un n-gramme requis par l'utilisateur. Cet outil fourni une liste de n-grammes ordonnés par corrélation descendante ainsi que la valeur de leurs coefficients de corrélation de Pearson. Il permet ensuite de naviguer à travers l'espace des corrélations fréquentielles entre les différents n-grammes par un simple clic sur l'un de ces n-grammes.

L'expérience de la recherche de mots par cet outil online à permis de mettre en évidence de nombreuses corrélations intéressantes. Ainsi, nous avons sélectionné un certain nombre de ces mots afin de représenter des thèmes variés. Nous présentons ensuite les résultats de recherche de corrélation de l'outil online sur ces mots en montrant les 20 premières corrélations pour chaque mot provenant de cet ensemble.

Nous présentons les vingt mots et n-grammes dont les profils fréquentiels sont les plus corrélés aux mots "amour", "bibliothèque", "coût", "couteau", "crise", "droit", "flammes", "guerre", "haine", "jeux", "juge", "littérature", "maladie", "paix", "philosophe", "politique", "professeur" et "spectacle" dans les Figures 14.10, 14.11 et 14.12.

amour		bibliothèque		coût	
aimer	0.884835	des sciences	0.708292	entreprises	0.925945
coeur	0.881416	la littérature	0.703913	prévu	0.922939
aime	0.856026	littérature	0.698617	l économie	0.919169
la jeune	0.837211	musée rath	0.685937	des entreprises	0.917832
regard	0.815779	professeur	0.676771	aux etats	0.916846
un regard	0.802453	rath	0.673778	alors que	0.916709
une femme	0.797332	des qualités	0.664695	aux etats unis	0.915376
bonheur	0.793489	le professeur	0.662393	économie	0.915039
le jeune	0.788553	tableaux	0.652624	millions de	0.909651
remords	0.787693	clarté	0.649831	aide	0.907577
sourire	0.786946	lullin	0.645156	cadre	0.907121
amoureux	0.785456	des travaux de	0.644018	notamment	0.906796
de bonheur	0.784512	arts	0.642774	alors que l	0.906203
amant	0.777934	simplicité	0.638403	possibilité de	0.905549
jalousie	0.776364	cette société	0.630721	base	0.904349
ma vie	0.773872	la liste des	0.630415	partiellement	0.903731
du ciel	0.772499	l excellent	0.629482	des produits	0.903725
de la jeune	0.770946	compositions	0.628938	la fin de l	0.903444
larmes	0.770652	l impression	0.624461	la réalité	0.903084
locales	0.769037	il fut	0.624334	lors des	0.901915
couteau		**crise**		**droit**	
cadavre	0.707279	réduction des	0.854938	donner à	0.904856
eu lieu dimanche	0.677154	financier	0.83293	autorité	0.901909
jeune	0.673623	une réduction	0.810051	soumettre	0.899566
le cadavre	0.662025	diminution	0.79485	il ne peut	0.892759
père	0.65125	diminution des	0.7857	qui ont	0.891726
mère	0.646453	la protection des	0.782395	opinion du	0.890671
fille	0.646059	de la situation	0.776817	des faits	0.889081
belle	0.642334	réduction	0.772073	les lois	0.888085
malade	0.642161	protection des	0.761592	des lois	0.887257
où a	0.635757	une diminution	0.757243	les droits	0.8859
doux	0.634138	financière	0.751424	il n y	0.884916
la victime	0.633785	manifestation	0.74749	opinion	0.884137
de dommages	0.632418	de réaliser	0.746369	laisser	0.883435
lieu dimanche	0.629293	de la politique	0.746273	principe	0.882935
soir un	0.627876	la diminution	0.746203	serait	0.88093
garçon	0.627572	programme	0.743613	a voulu	0.880152
en tous	0.627445	en vue d	0.743526	donner	0.879653
jeune fille	0.626529	organisée	0.742807	n a	0.879186
femme	0.626019	en vigueur	0.741946	qu il n y	0.878833
banquet	0.625223	économique	0.7416	n y	0.878317

TABLE 14.10 – Les 20 n-grammes les plus corrélés aux mots "amour", "bibliothèque", "coût", "couteau", "crise" et "droit"

flammes		guerre		haine	
un incendie	0.817325	allemands	0.89818	mépris	0.849657
feu a	0.774117	communique	0.883615	honte	0.837829
brisée	0.769014	allemands et	0.875467	accuser	0.832909
incendie a	0.767578	allemands qui	0.86704	courage	0.823672
éclata	0.761746	bataille	0.857435	le courage	0.816394
détruite	0.761085	une paix	0.857404	paisible	0.8132
le feu a	0.752386	attaque	0.856723	le malheur	0.810855
se fit	0.751312	champs de	0.855506	misérable	0.808207
dirigea	0.750671	armées	0.852101	odieux	0.806673
entendit	0.74584	commandement	0.848714	malheur	0.806116
passa	0.743144	une attaque	0.848331	indigne	0.804653
est tombée	0.742359	offensive	0.839104	tous les hommes	0.802366
ont été	0.739599	de soldats	0.837961	jamais	0.801833
donna	0.733876	la bataille	0.837883	des hommes qui	0.801642
en feu	0.733509	de munitions	0.835251	indignation	0.800662
rendit	0.733472	de l armée	0.834551	passions	0.799167
quitta	0.732916	les attaques	0.833317	la cause	0.798788
célébré	0.732909	l attaque	0.833152	conviction	0.798545
resta	0.732867	ennemie	0.832121	milieu	0.797085
la soirée	0.730009	attaques	0.829784	quelques hommes	0.794697
jeux		**juge**		**littérature**	
vainqueur	0.826363	d instruction	0.787978	directeurs	0.870889
finale	0.823185	justice	0.742644	littéraires	0.869197
lors des	0.791852	confiance	0.71469	les directeurs	0.858954
an dernier	0.790462	les causes	0.714252	de philosophie	0.85732
coupe	0.790407	ordre et la	0.708516	auteur	0.854891
l an dernier	0.790239	atteinte	0.706248	que l auteur	0.853718
champion	0.7901	prendre des mesures	0.705565	postes et	0.850673
million de	0.789331	l ordre et	0.704657	utiles	0.849206
pour la première fois	0.787304	de simple	0.703652	un genre	0.84477
la fin du	0.784608	l ordre et la	0.702548	des auteurs	0.840949
final	0.783345	ne peut	0.698041	modestie	0.839271
grand prix	0.78133	ceux qui ont	0.697051	l auteur	0.831491
classement	0.780258	tribunal	0.695693	le détail	0.830825
alors que	0.779539	juges	0.695434	défauts	0.828779
la première fois	0.777267	justice de	0.690904	pénitentiaire	0.821771
relève	0.777205	pas un	0.6909	arts	0.821297
notamment	0.776871	un coup	0.68989	auteurs	0.821019
pour la première	0.776558	lecture d	0.689166	philosophie	0.81998
alors que l	0.774059	des juges	0.689054	émulation	0.816218
aux etats unis	0.771987	sévir	0.686896	établir une	0.813651

TABLE 14.11 – Les 20 n-grammes les plus corrélés aux mots "flammes", "guerre", "haine", "jeux", "juge" et "littérature"

maladie		paix		philosophe	
le malade	0.690578	la guerre et	0.733445	la littérature	0.731045
après une longue	0.679729	guerre et	0.721242	littérature	0.710593
santé	0.667466	durable	0.702073	un siècle	0.686659
malade	0.667351	pendant la	0.687607	paysage	0.655585
est mort	0.665705	la guerre	0.684365	goût de	0.649033
l estomac	0.656955	de la victoire	0.675841	poètes	0.647346
après une courte	0.648948	l allemagne	0.674555	entre elles	0.641031
mutuels	0.647095	la guerre a	0.669984	prose	0.637797
parts	0.646852	apprend que	0.66858	style	0.633752
élu	0.646561	sacrifices	0.668505	science	0.630014
sanitaire	0.646117	de l allemagne	0.666678	siècle	0.629967
le minimum	0.645356	des intérêts	0.664902	arts	0.624466
a été très	0.644849	guerre a	0.664817	modestie	0.61883
est formé	0.644533	le point de vue	0.661057	l histoire	0.615214
arrivées	0.640954	du gouvernement	0.656628	charme	0.615013
après une	0.639486	le gouvernement des	0.656308	bibliothèque	0.614321
huit heures	0.638778	guerre	0.655378	petites	0.611884
et très	0.637511	gouvernement des	0.654852	la peinture	0.609548
une courte	0.633681	masses	0.652129	humain	0.609497
les médecins	0.633406	guerre il	0.649707	héros	0.608981

politique		professeur		spectacle	
des affaires	0.830671	recteur	0.822561	théâtre	0.766828
parti	0.802075	intéressant	0.794979	lancé	0.744244
avec les	0.7997	méthode	0.779353	scène	0.741598
des partis	0.798611	le succès	0.778052	jean	0.740916
à l égard	0.790236	clarté	0.776745	au cœur	0.738829
régime	0.786724	ensuite	0.77261	nuit	0.738331
vigueur	0.786068	la méthode	0.77248	mais aussi	0.717665
volonté	0.783971	l impression	0.767011	surprise	0.717159
le parti	0.782403	élèves	0.758542	en suisse	0.714952
politiques et	0.780412	approfondie	0.756199	jeu de	0.713335
le point	0.777923	auditeurs	0.753677	lance	0.711632
à l égard de	0.77706	obtenus	0.748752	pointe	0.69855
organisation	0.772556	il fut	0.744119	page	0.698284
à l égard du	0.771731	expose	0.738542	désormais	0.693844
l égard	0.770793	et demi	0.734694	bernard	0.693252
du parti	0.769643	convient	0.733952	réalisé	0.693015
l égard de	0.768773	impression	0.733796	une année	0.692964
propos	0.767772	succès de	0.731275	les années	0.692864
l égard du	0.767348	membres du comité	0.730423	le théâtre	0.690854
réaliser	0.766888	diverses	0.727509	culture	0.69036

TABLE 14.12 – Les 20 n-grammes les plus corrélés aux mots "maladie", "paix", "philosophe", "politique", "professeur" et "spectacle"

Nous observons sur ces exemples que la corrélation sur des fréquences relatives annuelles permet de mettre en évidence des n-grammes sémantiquement liés. La corrélation ne permet pas d'extraire un lien de causalité direct entre ces entités, mais l'objectif n'est pas de trouver ce lien de causalité qui est probablement le résultat d'une somme d'effets linguistiques complexes potentiellement liés à l'évolution de la sémantique des n-grammes. Globalement, nous observons qu'un mot est plus facilement lié à un autre mot plutôt qu'à des n-grammes de niveaux n supérieurs.

L'exemple du mot "amour" met en évidence premièrement le mot "aimer", ces deux entités ayant une racine similaire. Le deuxième mot, "coeur", ne possède cette fois aucune racine commune, mais est associé métaphoriquement au concept de l'amour (on aime avec le coeur plus qu'avec son cerveau). En continuant sur cet exemple, nous constatons que le mot "regard" est a priori plus éloigné de "amour" sémantiquement. Pourtant, il est extrêmement aisé pour un être humain de concevoir un rapprochement entre ces deux mots non pas sur la base du sens littéral des mots "amour" et "regard", mais parce que le sens commun nous enseigne la vision de l'amour romantique ou celle du coup de foudre comme étant un amour qui naît au premier regard. Cette hypothèse que nous proposons justifie alors la place du mot "regard" parmi les plus proches de "amour". Nous notons aussi la présence du mot "jalousie" correspondant à un effet souvent indésirable de l'amour qui est une émotion non rationnelle et considérée comme toxique conduisant au "coeur brisé" ou à la tristesse, raison probable pour laquelle les mots "larmes" et "remords" sont eux aussi présents dans la liste des vingt n-grammes les plus proches.

L'exemple du mot "bibliothèque" permet de constater un lien avec les sciences, la littérature, les musées, les arts et des notions intellectuelles ("professeur", "clarté", "simplicité") qui font penser à la compréhension et l'enseignement des savoirs. L'exemple du mot "coût" montre que l'utilisation de ce mot dans la presse est liée aux entreprises, aux prévisions budgétaires et à l'économie en générale. L'exemple du mot "couteau" est celui d'un ustensile de cuisine ou bien d'une arme. Dans le corpus de JDG, nous constatons que la présence du mot "banquet" est plutôt liée au premier sens. Le deuxième sens, plus présent dans ce corpus, est lié aux mots "cadavre", "victime" et "dommages". Etonnamment, nous retrouvons tous les mots décrivant les membres type de la famille proche "père", "mère", "fille", "garçon", etc. Cela fait probablement un lien entre les deux significations, car le couteau de cuisine est un instrument présent dans toutes les familles et est facilement accessible en cas de perte de contrôle émotionnel comme une arme facilement accessible dans le cadre d'une dispute. Bien entendu, ce sont des hypothèses fortes qui ne sont pas vérifiables de façon formelle et mathématique hormis une fouille qualitative dans les journaux au travers des outils mis à disposition. Plus étonnant encore est la notion temporelle des mots "soir" et "dimanche" où plusieurs hypothèses, à vérifier également, peuvent être imaginées du type "il y a plus de meurtres impliquant un couteau le dimanche et le soir", moment ou l'on se retrouve souvent en famille ou bien lors de sortie en boîte où l'alcool souvent présent augmente les réponses émotionnelles et donc ce type de risque.

L'exemple du mot "crise" met en évidence les mots "réduction", "diminution" et "protection" associés aux concepts décrits par "financier", "politique" et "économique". Ces mots permettent d'appréhender le sens du mot "crise" ou tout du moins les conséquences qui en découlent généralement. L'exemple du mot "droit" fait référence à des mots liés à sa sémantique comme "donner" (venant éventuellement de "donner le droit de"), "autorité", "soumettre", "faits" (exibant le 2-gramme "des faits" ce qui précise doublement qu'il ne s'agit pas de la conjugaison du verbe faire), "lois", "principe", etc. L'exemple du mot "flammes" montre une corrélation forte avec "incendie" (du 2-gramme "un incendie") ainsi que la façon d'annoncer cet événement particulier par le mot "éclate". En effet, dans l'usage de la langue, le sens commun impose souvent l'utilisation du verbe "éclate" et non pas "apparaît" par exemple. Les conséquences d'un tel événement sont également présentes par les corrélations avec les mots "brisée", "détruite", "tombée", "en feu". Les incendies sont des événements apparaissant plutôt aléatoirement et il semble qu'à une échelle annuelle les fréquences sont suffisamment différentes par année pour distinguer ces notions par une corrélation fréquentielle.

L'exemple du mot "guerre" est typique dans ce corpus, car nous avons déjà vu l'impact important des deux guerres mondiales dans les deux journaux GDL et JDG. C'est pour cela que dans ce cas l'histoire permet d'ajouter une corrélation importante entre les mots "guerre" et allemands". Les mots "bataille", "paix" (antonyme de "guerre"), "attaque", "armées", "commandement", "offensive", "soldats", "munitions" et "ennemie" ne souffrent d'aucune ambiguïté sémantique quant à leur place parmi les mots les plus corrélés avec "guerre". Nous informons également que "guerre" est corrélé à "état major" avec une valeur de 0.77 alors que sa corrélation avec "état" est de 0.13 tandis qu'avec "major" est de 0.69 (car "major" est quasiment toujours utilisé avec "état major") montrant l'intérêt de l'ajout de n-grammes de niveau supérieur dans certains cas particuliers. L'exemple du mot "haine", antonyme de "amour", met en évidence premièrement le mot "mépris" ce qui n'est pas anormal sachant que la haine qu'une personne éprouve via-à-vis d'une autre personne se manifeste souvent par du mépris. Ensuite les mots "honte" et "accuser" occupent les deuxièmes et troisièmes places, traduisant probablement le coté malveillant de ce mot et plus encore quand il s'agit de haine vis-à-vis d'une catégorie d'êtres humains causée notamment par des préjugés surtout compte tenu de notre histoire. Du reste, nous observons que la plupart des mots mis en évidence sont associés à un coté négatif comme "malheur", "misérable", "odieux", indigne", etc. Cela nous amène à faire l'hypothèse que le biais de positivité / négativité n'est pas constant dans le temps et que son évolution permet de distinguer les mots vus comme positifs (par exemple ceux associés à l'amour) de ceux qui sont vus comme négatifs (par exemple ceux associés à la haine).

L'exemple du mot "jeux" montre également des mots sémantiquement associés comme "vainqueur", "finale", "coupe", "champion", "final", "grand prix" et "classement". Il est étonnant de retrouver le 4-gramme "pour la première fois" qui pourrait marquer l'entrée dans l'histoire de joueurs qui furent inconnus avant ledit jeu. Il est à noter que l'importance des différents jeux de type "football", "hockey", "tennis", etc, s'est particulièrement accrue dans les vingt années les plus récentes du corpus. L'exemple du mot "juge" permet de comprendre rapidement que ce mot est souvent utilisé dans le cadre du 3-gramme "juge d'instruction". Le mot "confiance"

en deuxième position montre qu'il s'agit d'un élément important dans le domaine de la justice (confiance en la justice). Nous trouvons ensuite les mots "causes", "ordre", "atteinte", "tribunal", "justice" qui ont une liaison sémantique évidente. Nous observons également le 3-gramme "prendre des mesures" qui est l'un des rôles du juge afin que le délit ou crime ne se reproduise pas et que le prévenu puisse se réintégrer sauf cas exceptionnel. L'exemple du mot "littérature" montre les mots "littéraire" (avec la même racine), "philosophie, "auteur" et "genre" qui sont sémantiquement liés. D'autres mots ont une position plus difficile à décrypter dans cet espace de corrélation comme "directeurs", modestie", "pénitentiaire". Il est clair que certaines causes de ces corrélations sont loin d'être directes et celles-ci peuvent être étudiées plus précisément via le visualisateur de n-grammes.

L'exemple du mot "maladie" est aussi intéressant, car il fait le lien sémantique avec les mots "santé", "mort", "estomac", "sanitaire" et "médecins". Nous observons aussi les 3-grammes "après une longue" et "après une courte" qui sont deux options généralement utilisées pour décrire la durée de la maladie ("après une longue maladie" et "après une courte maladie"). L'exemple du mot "paix" fait le lien direct avec son antonyme "guerre" ainsi qu'avec "durable" (qui est généralement un objectif majeur concernant la paix dans le monde et entre les différents pays en général), "victoire", "allemagne", "sacrifice" et "gouvernement". Le mot "paix" est donc lié de façon importante au mot "guerre" dans ce corpus. L'exemple du mot "philosophe" met en évidence les mots "littérature", "poètes", "prose", "style", "science", "arts", "histoire", "bibliothèque", "peinture", "humain", etc qui sont des thèmes qui peuvent être associés à la philosophie. L'exemple du mot "politique" met également en évidence les mots liés au concept général comme "affaires", "parti", "régime", "vigueur", "volonté", "organisation", etc. L'exemple du mot "professeur" montre des mots qui sont liés à l'enseignement et l'école comme "recteur", "méthode", "clarté", "élèves", "approfondie", "auditeur", etc. L'exemple du mot "spectacle" montre des mots qui sont liés au théâtre et à la culture en général comme "théâtre", "scène", "surprise", "jeu", "culture", etc.

Cette expérience révèle que dans l'espace des profils fréquentiels annuels réside une information sémantique et plus particulièrement dans la comparaison de ceux-ci. D'autres méthodologies comme l'analyse des co-occurrences (Doddington, 2002; Lin et Hovy, 2003) permettent également de retrouver ce type d'informations. Ces méthodes utilisent généralement des fenêtres comme un nombre fixe de mots ou simplement un document. Une réduction de la granularité de la mesure fréquentielle à un article entier, à une journée (donc l'ensemble des articles d'une parution) ou à un mois permet probablement de retrouver des résultats similaires voir des meilleurs vu que la précision du calcul de la corrélation s'améliore avec le nombre de points.

Cependant, nos observations nous conduisent à valider l'hypothèse qu'à une échelle annuelle comportant au maximum 200 fréquences relatives, les profils fréquentiels des n-grammes contiennent une information sémantique dense.

Partie V
Conclusion et perspectives

Ce travail de thèse nous a permis d'approcher plusieurs objectifs dans le domaine de l'analyse quantitative de corpus diachroniques. Partant des travaux de Culturomics (Michel *et al.*, 2011), nous avons créé un visualisateur de n-grammes sur deux journaux "Gazette de Lausanne" (GDL) et "Journal de Genève" (JDG). Nous avons ensuite lié le visualisateur au site permettant la recherche directe dans les archives afin d'améliorer l'interopérabilité entre les outils. Cela donne la possibilité de passer d'une lecture distante à une lecture proche et d'explorer les journaux en utilisant le visualisateur de n-grammes comme point d'entrée. Cet outil, par analogie au microscope, nous a permis d'explorer une dimension quantitative à un niveau local (que nous avons appelé Micro) en analysant les profils fréquentiels des n-grammes.

Nous avons également exploré une dimension quantitative à un niveau plus global (que nous avons appelé Macro) en définissant des distances sur l'espace des subdivisions du corpus selon l'année de parution des articles. A ce niveau, un problème difficile de l'analyse de corpus est l'impact sur les mesures proposées de l'évolution même de la taille du corpus. Nous avons commencé l'analyse par la définition classique de distance de Jaccard (Jaccard, 1901) (Jaccard, 1912). Toutefois, en plus d'être affectée par la taille du corpus, la distance de Jaccard présente une sensibilité importante aux diverses erreurs d'OCR ainsi qu'à des phénomènes non linguistiques comme l'apparition de sections boursières, horaires de bus et train, etc.

Afin de contourner ces biais et d'éliminer ou réduire les effets d'évolution de taille de corpus ainsi que des phénomènes non liés à l'évolution linguistique, nous avons développé les concepts de noyau résilient et de n-grammes résilients. Ceux-ci nous ont permis d'adopter une vision ensembliste des mots et n-grammes selon leurs caractéristiques de résilience qui déterminent leur degré de stabilité dans le corpus. La notion de noyau résilient nous a également permis de cibler plus particulièrement l'évolution de la langue en excluant celle liée à la diversité des sujets, des articles ou des événements historiques. Nous avons ensuite proposé une nouvelle mesure basée sur ces concepts, la distance nucléaire. Bien que celle-ci soit fondamentalement différente de la distance de Jaccard, nous avons constaté que les deux mesures avaient un comportement proche.

Nous avons ensuite effectué plusieurs simulations afin d'étudier l'effet d'échantillonnage lié à l'évolution de la taille du corpus, considérant les deux corpus comme des échantillons de la langue dont nous souhaitons mesurer les évolutions. De façon attendue, la distance de Jaccard donne des résultats fluctuant selon la taille relative de l'échantillon par rapport au lexique simulé de la langue.

Cependant, nous avons constaté que la distance nucléaire est elle aussi affectée par les disparités de taille de corpus, tout en étant indépendante de la taille simulée du lexique de la langue. Cette propriété s'est révélée systématique quelle que soit la taille du lexique et ce sur une large gamme de tailles testées. Dès lors, il est possible d'estimer l'effet d'échantillonnage de la distance nucléaire et nous avons proposé un modèle linéaire simple qui vise à corriger la distance nucléaire réelle en lui retirant la valeur de l'effet d'échantillonnage par une opération de soustraction entre les distances calculées sur les données réelles et les distances simulées.

Nous avons également proposé une autre mesure exploitant la notion de noyau résilient. Il s'agit de l'entropie nucléaire, mesure d'entropie limitée à l'ensemble des éléments communs aux noyaux de JDG et GDL, permettant de comparer l'évolution des deux journaux. Des simulations d'effet d'échantillonnage ont permis de montrer que cette mesure n'est pas sensiblement affectée par les variations de taille de corpus, si ce n'est quand la taille est trop faible (ce qui est le cas des années antérieures à 1840) et donc que l'échantillon n'est pas suffisamment représentatif.

Nous avons aussi créé une représentation visuelle, le chronocloud, permettant de visualiser l'ensemble des n-grammes d'un corpus et leurs caractéristiques sous forme de nuages structurés de n-grammes. Cette visualisation se situe entre les niveaux Macro et Micro, faisant le lien entre une vision distante de l'ensemble du corpus et le profil fréquentiel des n-grammes individuels qui le composent. Ainsi, il est possible de zoomer rapidement (images pré-calculées via la technologie deepzoom) sur les n-grammes intéressants, positionnés selon les caractéristiques de fréquence moyenne, de résilience et d'année de fréquence maximale pour ensuite accéder à leurs profils fréquentiels et éventuellement à une recherche directe dans les archives.

Nous avons appliqué ces outils et mesures aux deux corpus JDG et GDL tout en explorant l'échelle n des n-grammes. Cela nous a permis d'observer des évolutions intéressantes tant au niveau Micro que Macro. Au niveau Macro, nous avons observé sur les vingt dernières années du corpus un phénomène global de diminution de l'entropie nucléaire due à divers changements accélérés dans les profils fréquentiels des n-grammes appartenant au noyau résilient. Au niveau Micro, nous avons observé notamment des effets comme ceux de la réforme orthographique de 1835 ou les remplacements de n-grammes par d'autres dont la signification est similaire. Nous avons pu observer la durée de ces phénomènes de remplacement (généralement une dizaine d'années) et des différences de temporalité selon le journal.

Nous avons également défini des mesures agrégeant les différents niveaux n des n-grammes selon des pondérations définies sur la base de nos observations concernant les contributions informationnelles de ces différents niveaux. Nous avons aussi créé une visualisation chronocloud globale définie sur tous les niveaux n des n-grammes. Les relations fondamentales entre n-grammes inclus les uns dans les autres ont été explorées afin d'établir une méthode de décomposition des profils fréquentiels, d'analyser l'histoire d'un n-gramme au travers de sa décomposition minimale et de détecter les expressions solidifiées. En outre, il a été possible de distinguer des changements sémantiques sur la base de décompositions minimales, comme par exemple celui du mot "relativement".

Nous avons aussi exploré les relations qu'entretiennent les n-grammes entre-eux, quand ils ne sont pas liés par l'équation de décomposition, via l'utilisation du coefficient de corrélation de Pearson entre leurs profils fréquentiels. Cela nous a permis d'extraire de l'information sémantique sur la base des courbes de profils fréquentiels et donc de l'histoire fréquentielle annuelle du n-gramme.

Bien que plusieurs notions fondamentales aient été explorées au cours de cette thèse, il est clair que ce travail est un prélude ouvrant la voie à d'autres études approfondies pour chacun de ces concepts. Ceux-ci sont mathématiquement simples et parfois explorés ou formalisés différemment dans le cadre d'études existantes. Nous avons tenté de lier ces concepts entre eux afin de tirer autant d'informations linguistiques que possible sur les deux corpus de JDG et GDL aux niveaux Macro et Micro ainsi que selon les niveaux n des n-grammes. Les relations entre ces niveaux n des n-grammes, relativement peu explorées jusqu'ici, ont montré un potentiel intéressant pour toute étude de corpus diachronique.

Par conséquent de nombreux travaux futurs sont possibles. Les concepts de noyau et mots résilients sont par définition des notions ensemblistes et il est donc possible d'établir une algèbre des ensembles sur un corpus diachronique sur la base de ces notions. De plus, certains corpus particuliers comme celui de Google Books (Michel *et al.*, 2011) ont une taille plus large encore. Dans ce contexte, les n-grammes du noyau résilient et les profils fréquentiels sont plus facilement non nuls même s'ils sont proches de zéro. Il est alors possible d'imaginer une nouvelle approche combinant un filtre fréquentiel en sus du noyau résilient afin de mieux cibler les n-grammes plus importants pour étudier la langue. Toutefois, la mise en oeuvre de ces concepts sur les corpus de GDL et JDG ne requiert pas ce type d'adaptation.

La visualisation chronocloud peut également être améliorée selon plusieurs pistes. En premier lieu, le concept de polycloud est indépendant du temps et peut servir à afficher des n-grammes selon diverses caractéristiques désirées, ce qui permet d'explorer plusieurs dimensions différentes sans changer fondamentalement de méthode. Plus spécifiquement, le chronocloud possède des sections positionnées selon les caractéristiques désirées, mais les positions à l'intérieur de ces sections restent aléatoires et produites par un algorithme de classique de création de nuages de mots. Une voie d'amélioration serait de calculer directement les positions des n-grammes à l'intérieur des sections selon ces mêmes caractéristiques. Ce type d'algorithme conserverait une composante aléatoire, mais représenterait aussi les caractéristiques des n-grammes avec plus de précision.

Le principe même de la décomposition multi-échelle minimale des n-grammes autorise plusieurs façons de procéder. Nous avons utilisé une modélisation de type gaussienne, car nous avons observé qu'elle fonctionne dans un nombre important de cas et d'autant plus en progressant dans les niveaux n. Toutefois, le modèle choisi est arbitraire et nous savons que certains profils fréquentiels de n-grammes correspondent mieux à d'autres modélisations, car différents types de courbes ont été rencontrés dans cette étude. Par exemple, la modélisation gaussienne ne fonctionne pas pour les phénomènes asymétriques rattachés aux mots composés de caractères numériques correspondant aux années, qui ont une fréquence d'utilisation importante l'année en question, mais conservent une mémoire dans le temps, de telle façon que leur profil fréquentiel en devient asymétrique. Evidemment, en poussant le raisonnement mathématique à la limite, le profil fréquentiel de tout n-gramme finit par devenir un delta de Kronecker pour un niveau n suffisamment élevé, du moins si ceux-ci ne sont pas filtrés par une fréquence minimale.

L'espace des corrélations des profils fréquentiels des n-grammes ouvre également des possibilités d'étude importantes, car il n'est pas complètement intuitif qu'une simple corrélation entre deux profils fréquentiels puisse révéler des relations sémantiques complexes. En effet, nos observations montrent que les mots et n-grammes rattachés à un thème particulier, donnant lieu à d'importants changements fréquentiels, sont liés par la corrélation. En poussant le raisonnement jusqu'au bout, la méthode peut être vue comme une étude de co-occurrences dont la fenêtre de co-occurrence serait élargie aux données correspondant à une année entière. Dans notre cas, la fenêtre est variable en terme de nombre de mots, mais c'est la temporalité et donc l'histoire fréquentielle des mots et n-grammes qui permet de les rassembler, entre autre, en fonction de leur sémantique. Ainsi, le lien entre les méthodes de co-occurrence et de corrélation des profils fréquentiels mérite également une étude plus approfondie. Nous avons tenté d'explorer une diversité de méthodologies dans le but de cibler plus particulièrement les éléments qui ont un rôle important dans l'étude de la langue plutôt que les phénomènes sporadiques causés par les événements historiques ou des évolutions non linguistiques. Il a fallu pour cela proposer des méthodes robustes et innovantes ainsi que de nouvelles visualisations, afin de disposer d'un panel d'outils destinés à mesurer l'évolution linguistique des corpus. Ces méthodes se déclinent selon différents niveaux de proximité au texte allant de méthodes distantes jusqu'à des analyses de phénomènes "au microscope" dans un niveau plus proche du contenu textuel du corpus.

Nous avons exploré chaque niveau n des n-grammes indépendamment en utilisant nos outils sur les corpus de GDL et JDG, puis nous avons posé la question de l'unification de ces niveaux afin de concevoir des mesures et visualisations multi-échelle. En effet, nous avons observé que l'information contenue dans les niveaux supérieurs était également importante, et indissociable des autres niveaux. Ainsi, nous avons constaté que la relation entre ces niveaux n (le plus souvent étudiés séparément) permet de dégager un certain nombre de propriétés intéressantes des entités étudiées, comme le montre les résultats de la méthode de décomposition minimale des n-grammes. De plus, une méthode multi-échelle peut-être plus facilement adaptée à l'étude d'autres unités de la langue, comme par exemple les n-grammes de caractères. Ainsi, nous concluons qu'une plus grande attention devrait être portée à l'avenir sur l'étude multi-échelle des n-grammes et qu'elle devrait également se porter sur la relation qu'entretiennent les différents niveaux n d'analyse des n-grammes entre-eux. En outre, ces mêmes méthodes multi-échelles devraient idéalement être appliquées sur les n-grammes de caractères afin de rendre l'étude indépendante des contraintes et hypothèses nécessaires à la détermination des mots comme unité de base, par exemple la tokenisation.

Selon les mesures que nous avons effectué au niveau Macro, il n'a pas été possible de montrer que l'évolution de la langue, au travers de ces corpus de presse, est en phase d'accélération ou de décélération. L'évolution linguistique du corpus est présente mais stable. Toutefois, il a été possible de montrer une évolution au niveau de la distribution fréquentielle des n-grammes du noyau résilient. Ceci s'est reflété par la mesure d'entropie nucléaire qui baisse constamment à partir de 1950. De futurs travaux devraient effectuer le même type de mesures sur d'autres corpus afin de vérifier si cette évolution est propre à la langue et non pas au corpus.

La contribution de cette thèse est essentiellement méthodologique. Cette approche de type Big Data présente des atouts majeurs, permettant rapidement d'obtenir des résultats pertinents et intéressants pour diverses questions de recherche dont en particulier l'étude de l'évolution de la langue au travers d'un corpus donné. Ainsi, nous espérons avoir contribué au développement de méthodes automatiques et totalement indépendantes du langage dans lequel est écrit le corpus. Rappelons que nous considérons simplement une suite de signes, les n-grammes de mots. Nous avons ainsi volontairement minimisé les hypothèses de travail afin de garder une possibilité de généralisation importante des méthodes développées et permettre leur application à différents corpus avec peu ou pas d'adaptations méthodologiques. Ainsi, on pourrait également espérer obtenir des résultats sur un corpus dont on ne connait rien, pas même la langue dans lequel il a été écrit, un corpus très ancien et en poussant le raisonnement jusqu'au bout, un corpus de langue extraterrestre.

Toutefois, il est aussi avéré sur la base des travaux menés dans cette thèse sur les corpus journalistiques de JDG et GDL que la connaissance du corpus et de la langue sont évidemment des atouts importants. En effet, même si la plupart des outils que nous avons développés fournissent des résultats automatiques, la portée des hypothèses qui sous-tendent les analyses effectuées n'en serait que multipliée si ces méthodes étaient combinées avec l'analyse qualitative d'un linguiste ayant une solide expérience. La complémentarité est donc essentielle sur ces deux aspects, bien que ce travail de thèse se préoccupe principalement de méthodes d'analyse automatique générant des résultats dans le cadre de grands corpus textuels diachroniques.

Ces méthodes doivent donc être vues comme une aide pour le linguiste permettant d'identifier automatiquement des phénomènes potentiellement intéressants et de faciliter le travail d'analyse. Ainsi, ces méthodes ont pour objectif de permettre au linguiste de gagner un temps précieux en raison de leur caractère général, indépendant de la langue et du faible nombre d'hypothèses requises pour les appliquer.

Enfin, il est important que ces méthodes soient testées sur d'autres corpus de presse afin de permettre la comparaison de leur évolution linguistique, mais également sur des corpus d'autres natures. En effet, les méthodes développées dans ce travail ont été conçues dans une perspective de robustesse vis-à-vis d'artefacts qui ne sont pas dus à des évolutions linguistiques. Cela peut être dû à des évolutions de sujets, à des résultats de l'OCR qui sont fluctuants selon les époques du corpus (qualité de la numérisation, évolution de la mise en page, etc.) ou simplement à l'évolution de la taille du corpus. Des corpus d'autres natures pourraient révéler des effets qui n'ont pas été détectés dans les corpus de journaux.

Ce domaine, popularisé par les travaux "culturomics" (Michel *et al.*, 2011), est en plein développement et permet d'étudier la langue au travers des corpus textuels. Il est nécessaire de continuer à proposer des méthodologies quantitatives robustes afin d'exploiter le potentiel de l'analyse de ces gigadonnées, d'aider les chercheurs, historiens, sociologues ou encore les linguistes à gagner un temps précieux, de faciliter la découverte de nouveaux savoirs sur l'évolution linguistique du passé pour éclairer les évolutions présentes ou même futures.

Bibliographie

Tony ABOU-ASSALEH, Nick CERCONE, Vlado KESELJ et Ray SWEIDAN : Detection of new malicious code using n-grams signatures. *In PST*, pages 193–196, 2004.

Chris ANDERSON : The end of theory : The data deluge makes the scientific method obsolete. *Wired magazine*, 16(7):16–07, 2008.

Jacques ARENDS, Pieter MUYSKEN et Norval SMITH : *Pidgins and creoles : An introduction*, volume 15. John Benjamins Publishing, 1994.

Farzindar ATEFEH et Wael KHREICH : A survey of techniques for event detection in twitter. *Computational Intelligence*, 31(1):132–164, 2015.

Quentin D ATKINSON et Russell D GRAY : Curious parallels and curious connections—phylogenetic thinking in biology and historical linguistics. *Systematic biology*, 54 (4):513–526, 2005.

Julia BAMFORD, Silvia CAVALIERI et Giuliana DIANI : *Variation and change in spoken and written discourse : Perspectives from corpus linguistics*, volume 21. 2013. ISBN 9789027210388.

Michele BANKO et Lucy VANDERWENDE : Using n-grams to understand the nature of summaries. *In Proceedings of HLT-NAACL 2004 : Short Papers*, pages 1–4. Association for Computational Linguistics, 2004.

Alberto BARRÓN-CEDEÑO et Paolo ROSSO : On automatic plagiarism detection based on n-grams comparison. *Advances in Information Retrieval*, pages 696–700, 2009.

Mark BARTLETT et Dimitar KAZAKOV : The origins of syntax : from navigation to language. *Connection Science*, 17(3-4):271–288, 2005.

Youssef BASSIL et Mohammad ALWANI : Ocr context-sensitive error correction based on google web 1t 5-gram data set. *arXiv preprint arXiv :1204.0188*, 2012a.

Youssef BASSIL et Mohammad ALWANI : Ocr post-processing error correction algorithm using google online spelling suggestion. *arXiv preprint arXiv :1204.0191*, 2012b.

Emily BENDER et Jeff GOOD : A grand challenge for linguistics : Scaling up and integrating models. *White paper contributed to NSF's SBE*, 2020:1–1, 2010.

Frank BENFORD : The law of anomalous numbers. *Proceedings of the American philosophical society*, pages 551–572, 1938.

Vinay BETTADAPURA, Grant SCHINDLER, Thomas PLÖTZ et Irfan ESSA : Augmenting bag-of-words : Data-driven discovery of temporal and structural information for activity recognition. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2619–2626, 2013.

Douglas BIBER : Representativeness in corpus design. *Literary and linguistic computing*, 8 (4):243–257, 1993.

Ismaïl BISKRI, Jean-Guy MEUNIER et Sylvain JOYAL : L'extraction des termes complexes : une approche modulaire semiautomatique. *Presses Universitaires de Louvain*, 1:192201, 2004.

David M BLEI et John D LAFFERTY : Visualizing topics with multi-word expressions. *arXiv preprint arXiv :0907.1013*, 2009.

V. BOCHKAREV, V. SOLOVYEV et S. WICHMANN : Universals versus historical contingencies in lexical evolution. *Journal of The Royal Society Interface*, 11(101), 2014. ISSN 1742-5689.

George EP BOX : Science and statistics. *Journal of the American Statistical Association*, 71 (356):791–799, 1976.

Etienne BRUNET : Peut-on mesurer la distance entre deux textes ? *Corpus*, (2), 2003.

Marc BRYSBAERT, Matthias BUCHMEIER, Markus CONRAD, Arthur M JACOBS, Jens BÖLTE et Andrea BÖHL : The word frequency effect. *Experimental psychology*, 2011.

Christian BUCK, Kenneth HEAFIELD et Bas VAN OOYEN : N-gram counts and language models from the common crawl. *In LREC*, volume 2, page 4, 2014.

Vincent BUNTINX, Cyril BORNET et Frédéric KAPLAN : Studying Linguistic Changes on 200 Years of Newspapers. *In Digital Humanities 2016*, 2016.

Vincent BUNTINX, Cyril BORNET et Frédéric KAPLAN : Studying linguistic changes over 200 years of newspapers through resilient words analysis. *Frontiers in Digital Humanities*, 4:2, 2017a. ISSN 2297-2668. URL http://journal.frontiersin.org/article/10.3389/fdigh.2017.00002.

Vincent BUNTINX et Frédéric KAPLAN : Inversed N-gram viewer : Searching the space of word temporal profiles. *In Digital Humanities 2015*, 2015.

Vincent BUNTINX, Aris XANTHOS et Frédéric KAPLAN : Layout Analysis on Newspaper Archives. *In Digital Humanities 2017*, 2017b.

M. BURCH, S. LOHMANN, F. BECK, N. RODRIGUEZ, L. D. SILVESTRO et D. WEISKOPF : Radcloud : Visualizing multiple texts with merged word clouds. *In 2014 18th International Conference on Information Visualisation*, pages 108–113, July 2014.

Nicholas CARR : *The shallows : What the Internet is doing to our brains*. WW Norton & Company, 2011.

Quim CASTELLÀ et Charles SUTTON : Word storms : Multiples of word clouds for visual comparison of documents. *In Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 665–676, New York, NY, USA, 2014. ACM.

Quim CASTELLÀ et Charles A. SUTTON : Word storms : Multiples of word clouds for visual comparison of documents. *CoRR*, abs/1301.0503, 2013.

William B CAVNAR, John M TRENKLE *et al.* : N-gram-based text categorization. *Ann Arbor MI*, 48113(2):161–175, 1994.

Nathanael CHAMBERS : Labeling documents with timestamps : Learning from their time expressions. *In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics : Long Papers-Volume 1*, pages 98–106. Association for Computational Linguistics, 2012.

David CHAVALARIAS et Jean-Philippe COINTET : Phylomemetic patterns in science evolution—the rise and fall of scientific fields. *PloS one*, 8(2):e54847, 2013.

Himani CHAWLA : Facebook page spam detection using support vector machines based on n-gram model. *International Journal of Computer Science Issues (IJCSI)*, 11(5):161, 2014.

Costin-Gabriel CHIRU et Madalina TOIA : Using time series analysis for estimating the time stamp of a text. *In International Work-Conference on Time Series Analysis*, pages 35–47. Springer, 2016.

Kyunghyun CHO, Bart VAN MERRIËNBOER, Caglar GULCEHRE, Dzmitry BAHDANAU, Fethi BOUGARES, Holger SCHWENK et Yoshua BENGIO : Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv :1406.1078*, 2014.

Morten H CHRISTIANSEN et Simon KIRBY : Language evolution : Consensus and controversies. *Trends in cognitive sciences*, 7(7):300–307, 2003.

Ali ÇILTIK et Tunga GÜNGÖR : Time-efficient spam e-mail filtering using n-gram models. *Pattern Recognition Letters*, 29(1):19–33, 2008.

Rudolf CLAUSIUS : *Théorie mécanique de la chaleur*, volume 2. Eugène Lacroix, 1868.

Germinal COCHO, Jorge FLORES, Carlos GERSHENSON, Carlos PINEDA et Sergio SÁNCHEZ : Rank diversity of languages : Generic behavior in computational linguistics. *PLOS ONE*, 10 (4):1–12, 04 2015.

C. COLLINS, F. B. VIEGAS et M. WATTENBERG : Parallel tag clouds to explore and analyze faceted text corpora. *In 2009 IEEE Symposium on Visual Analytics Science and Technology*, pages 91–98, Oct 2009.

Matthieu CONSTANT, Anthony SIGOGNE et Patrick WATRIN : Discriminative strategies to integrate multiword expression recognition and parsing. *In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics : Long Papers-Volume 1*, pages 204–212. Association for Computational Linguistics, 2012.

Glen COPPERSMITH et Erin KELLY : Dynamic wordclouds and vennclouds for exploratory data analysis. *In Association for Computational Linguistics Workshop on Interactive Language Learning and Visualization*, 2014.

Michael CRAWFORD, Taghi M KHOSHGOFTAAR, Joseph D PRUSA, Aaron N RICHTER et Hamzah AL NAJADA : Survey of review spam detection using machine learning techniques. *Journal of Big Data*, 2(1):23, 2015.

Christopher CULY et Susanne Z RIEHEMANN : The limits of n-gram translation evaluation metrics. *In MT Summit IX*, pages 71–78, 2003.

Marc DAMASHEK : Gauging similarity with n-grams : Language-independent categorization of text. *Science*, 267(5199):843, 1995.

Jean-Paul DELAHAYE et Nicolas GAUVRIT : *Culturomics : le numérique et la culture*. Odile Jacob, 2013.

Jiandong DING, Shuigeng ZHOU et Jihong GUAN : mirfam : an effective automatic mirna classification method based on n-grams and a multiclass svm. *BMC bioinformatics*, 12 (1):216, 2011.

Stefanie DIPPER : Theory-driven and corpus-driven computational linguistics and the use of corpora. 2008.

George DODDINGTON : Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *In Proceedings of the second international conference on Human Language Technology Research*, pages 138–145. Morgan Kaufmann Publishers Inc., 2002.

Nadir DURRANI, Helmut SCHMID, Alexander FRASER, Philipp KOEHN et Hinrich SCHÜTZE : The operation sequence model—combining n-gram-based and phrase-based statistical machine translation. *Computational Linguistics*, 2015.

Leo EGGHE : Untangling herdan's law and heaps' law : Mathematical and informetric arguments. *Journal of the American Society for Information Science and Technology*, 58 (5):702–709, 2007.

Maud EHRMANN, Giovanni COLAVIZZA, Yannick ROCHAT et Frédéric KAPLAN : Diachronic evaluation of ner systems on old newspapers. *In Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, numéro EPFL-CONF-221391, pages 97–107, 2016.

Martin EMMS et Arun JAYAPAL : Detecting change and emergence for multiword expressions. *In MWE@ EACL*, pages 89–93, 2014.

Günes ERKAN et Dragomir R RADEV : Lexrank : Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479, 2004.

John EVERSHED et Kent FITCH : Correcting noisy ocr : Context beats confusion. *In Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, pages 45–51. ACM, 2014.

Marcello FEDERICO et Mauro CETTOLO : Efficient handling of n-gram language models for statistical machine translation. *In Proceedings of the Second Workshop on Statistical Machine Translation*, pages 88–95. Association for Computational Linguistics, 2007.

Udo FRIES et Hans Martin LEHMANN : The style of 18th century english newspapers : Lexical diversity. *News Discourse in Early Modern Britain*, pages 91–104, 2006.

Francis GALTON : Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263, 1886.

Philippe GAMBETTE et Jean VÉRONIS : *Visualising a Text with a Tree Cloud*, pages 561–569. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

Madhavi GANAPATHIRAJU, Deborah WEISSER, Roni ROSENFELD, Jaime CARBONELL, Raj REDDY et Judith KLEIN-SEETHARAMAN : Comparative n-gram analysis of whole-genome protein sequences. *In Proceedings of the second international conference on Human Language Technology Research*, pages 76–81. Morgan Kaufmann Publishers Inc., 2002.

Jianbo GAO, Jing HU, Xiang MAO et Matjaž PERC : Culturomics meets random fractal theory : insights into long-range correlations of social and natural phenomena over the past two centuries. *Journal of The Royal Society Interface*, page rsif20110846, 2012.

Anne GARCIA-FERNANDEZ, Anne-Laure LIGOZAT, Marco DINARELLI et Delphine BERNHARD : Méthodes pour l'archéologie linguistique : datation par combinaison d'indices temporels. *Actes du septième DÉfi Fouille de Textes*, page 29, 2011.

Martin GERLACH, Francesc FONT-CLOS et Eduardo G. ALTMANN : Similarity of symbol frequency distributions with heavy tails. *Phys. Rev. X*, 6:021009, Apr 2016.

George GIANNAKOPOULOS, Vangelis KARKALETSIS, George VOUROS et Panagiotis STAMATOPOULOS : Summarization system evaluation revisited : N-gram graphs. *ACM Transactions on Speech and Language Processing (TSLP)*, 5(3):5, 2008.

Daniel GILDEA et Thomas HOFMANN : Topic-based language models using em. *In Sixth European Conference on Speech Communication and Technology*, 1999.

Thomas GOTTRON : *Document Word Clouds : Visualising Web Documents as Tag Clouds to Aid Users in Relevance Decisions*, pages 94–105. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

Casey S GREENE, Jie TAN, Matthew UNG, Jason H MOORE et Chao CHENG : Big data bioinformatics. *Journal of cellular physiology*, 229(12):1896–1900, 2014.

E GREEVY et S SMEATON : Text categorization of racist texts using a support vector machine. *7 es Journées internationales d'Analyse statistique des Données Textuelles*, 2004.

Thiago S GUZELLA et Walmir M CAMINHAS : A review of machine learning approaches to spam filtering. *Expert Systems with Applications*, 36(7):10206–10222, 2009.

Md Akmal HAIDAR et Douglas O'SHAUGHNESSY : Topic n-gram count language model adaptation for speech recognition. *In Spoken Language Technology Workshop (SLT), 2012 IEEE*, pages 165–169. IEEE, 2012.

Juha HAKKINEN et Jilei TIAN : N-gram and decision tree based language identification for written words. *In Automatic Speech Recognition and Understanding, 2001. ASRU'01. IEEE Workshop on*, pages 335–338. IEEE, 2001.

Raffay HAMID, Siddhartha MADDI, Aaron BOBICK et Irfan ESSA : Unsupervised analysis of activity sequences using event-motifs. *In Proceedings of the 4th ACM international workshop on Video surveillance and sensor networks*, pages 71–78. ACM, 2006.

William L HAMILTON, Jure LESKOVEC et Dan JURAFSKY : Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv :1605.09096*, 2016.

Hilda HARDY, Nobuyuki SHIMIZU, Tomek STRZALKOWSKI, Liu TING, Xinyang ZHANG et G Bowden WISE : Cross-document summarization by concept classification. *In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 121–128. ACM, 2002.

Y. HASSAN-MONTERO et V. HERRERO-SOLANA : Improving tag-clouds as visual information retrieval interfaces. *In Proc. InSciT 2006*, Merida, Spain, octobre 2006.

Harold Stanley HEAPS : *Information retrieval : Computational and theoretical aspects*. Academic Press, Inc., 1978.

F. HEIMERL, S. LOHMANN, S. LANGE et T. ERTL : Word cloud explorer : Text analytics based on word clouds. *In 2014 47th Hawaii International Conference on System Sciences*, pages 1833–1842, Jan 2014.

Gustav HERDAN : Quantitative linguistics. 1964.

Christopher J HOWE et Heather F WINDRAM : Phylomemetics—evolutionary analysis beyond the gene. *PLoS biology*, 9(5):e1001069, 2011.

Anette HULTH : Improved automatic keyword extraction given more linguistic knowledge. *In Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 216–223. Association for Computational Linguistics, 2003.

Paul JACCARD : Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37:547–579, 1901.

Paul JACCARD : The distribution of the flora in the alpine zone.1. *New Phytologist*, 11(2):37–50, 1912. ISSN 1469-8137.

Radwan JALAM et Jean-Hugues CHAUCHAT : Pourquoi les n-grammes permettent de classer des textes ? recherche de mots-clefs pertinents à l'aide des n-grammes caractéristiques. *In 6th International Conference on Textual Data Statistical Analysis, France*, pages 381–390, 2002.

Michèle JARDINO : Identification des auteurs de textes courts avec des n-grammes de caractères. *Proc. of JADT'2006 (8èmes journées internationales d'Analyse Statistique des Données Textuelles)*, 2:543–549, 2006.

Adam JATOWT et Kevin DUH : A framework for analyzing semantic change of words across time. *In Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 229–238. IEEE Press, 2014.

Guofei JIANG, Haifeng CHEN, Cristian UNGUREANU et Kenji YOSHIHIRA : Multiresolution abnormal trace detection using varied-length *n*-grams and automata. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 37(1):86–97, 2007.

Patrick JUOLA : Using the google n-gram corpus to measure cultural complexity. *Literary and linguistic computing*, 28(4):668–675, 2013.

Ioannis KANARIS, Konstantinos KANARIS, Ioannis HOUVARDAS et Efstathios STAMATATOS : Words versus character n-grams for anti-spam filtering. *International Journal on Artificial Intelligence Tools*, 16(06):1047–1067, 2007.

Frederic KAPLAN : Linguistic capitalism and algorithmic mediation. *Representations*, 127 (1):57–63, 2014.

Frédéric KAPLAN et Isabella di LENARDO : Big data of the past. *Frontiers in Digital Humanities*, 4:12, 2017.

Rianne KAPTEIN et Jaap KAMPS : *Word Clouds of Multiple Search Results*, pages 78–93. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

Owen KASER et Daniel LEMIRE : Tag-cloud drawing : Algorithms for cloud visualization. *CoRR*, abs/cs/0703109, 2007.

Jinyun KE et John H HOLLAND : Language origin from an emergentist perspective. *Applied Linguistics*, 27(4):691–716, 2006.

Su Nam KIM, Timothy BALDWIN et Min-Yen KAN : Evaluating n-gram based evaluation metrics for automatic keyphrase extraction. *In Proceedings of the 23rd international conference on computational linguistics*, pages 572–580. Association for Computational Linguistics, 2010.

Simon KIRBY : Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, 5 (2):102–110, 2001.

Pranam KOLARI, Akshay JAVA, Tim FININ, Tim OATES et Anupam JOSHI : Detecting spam blogs : A machine learning approach. *In AAAI*, volume 6, pages 1351–1356, 2006.

Moshe KOPPEL, Jonathan SCHLER et Shlomo ARGAMON : Computational methods in authorship attribution. *Journal of the Association for Information Science and Technology*, 60 (1):9–26, 2009.

Kosmas KOSMIDIS, John M HALLEY et Panos ARGYRAKIS : Language evolution and population dynamics in a system of two interacting species. *Physica A : Statistical Mechanics and its Applications*, 353:595–612, 2005.

Vivek KULKARNI, Rami AL-RFOU, Bryan PEROZZI et Steven SKIENA : Statistically significant detection of linguistic change. *In Proceedings of the 24th International Conference on World Wide Web*, pages 625–635. International World Wide Web Conferences Steering Committee, 2015.

S KULLBACK : Letters to the editor. *The American Statistician*, 41(4):338–341, 1987.

S. KULLBACK et R. A. LEIBLER : On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86, 03 1951.

Abhimanu KUMAR, Matthew LEASE et Jason BALDRIDGE : Supervised language modeling for temporal resolution of texts. *In Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 2069–2072. ACM, 2011.

Atul KUMAR : A survey on various ocr errors. *International Journal of Computer Applications*, 143(4):8–10, 2016.

William LABOV : *Principles of linguistic change. Vol. 1 : Internal features*. Oxford : Blackwell, 1994.

Vasileios LAMPOS et Nello CRISTIANINI : Nowcasting events from the social web with statistical learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(4):72, 2012.

B. LEE, N. H. RICHE, A. K. KARLSON et S. CARPENDALE : Sparkclouds : Visualizing trends in tag clouds. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1182–1189, Nov 2010.

Michael LEVANDOWSKY et David WINTER : Distance between sets. *Nature*, 234:34 – 35, 1971.

Yuanpeng LI, Dmitriy GENZEL, Yasuhisa FUJII et Ashok C POPAT : Publication date estimation for printed historical documents using convolutional neural networks. *In Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing*, pages 99–106. ACM, 2015.

Chin-Yew LIN : Rouge : A package for automatic evaluation of summaries. *In Text summarization branches out : Proceedings of the ACL-04 workshop*, volume 8. Barcelona, Spain, 2004.

Chin-Yew LIN et Eduard HOVY : Automatic evaluation of summaries using n-gram co-occurrence statistics. *In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 71–78. Association for Computational Linguistics, 2003.

Nikola LJUBESIC, Nives MIKELIC et Damir BORAS : Language indentification : How to distinguish similar languages ? *In Information Technology Interfaces, 2007. ITI 2007. 29th International Conference on*, pages 541–546. IEEE, 2007.

Gunn Inger LYSE et Gisle ANDERSEN : Collocations and statistical analysis of n-grams. *Exploring Newspaper Language : Using the Web to Create and Investigate a Large Corpus of Modern Norwegian,, Studies in Corpus Linguistics, John Benjamins Publishing, Amsterdam*, pages 79–109, 2012.

Benoit MANDELBROT : Information theory and psycholinguistics. *BB Wolman and E*, 1965.

Carla MARCEAU : Characterizing the behavior of a program using multiple-length n-grams. *In Proceedings of the 2000 workshop on New security paradigms*, pages 101–110. ACM, 2001.

Damon MAYAFFRE : Rôle et place du corpus en linguistique. réflexions introductives. *In Actes du colloque JETOU'2005*, pages 5–17. Université de Toulouse-Le Mirail, 2005.

April MS MCMAHON : *Understanding language change*. Cambridge University Press, 1994.

Paul MCNAMEE : N-gram tokenization for indian language text retrieval. *In Working Notes of the Forum for Information Retrieval Evaluation*, pages 12–14, 2008.

Paul MCNAMEE et James MAYFIELD : Character n-gram tokenization for european language text retrieval. *Information retrieval*, 7(1):73–97, 2004.

C MCNAUGHT et P LAM : Using wordle as a supplementary research tool. *The Qualitative Report*, 15(3):630–643, 2010.

Scott W MCQUIGGAN, Sunyoung LEE et James C LESTER : Early prediction of student frustration. *In International Conference on Affective Computing and Intelligent Interaction*, pages 698–709. Springer, 2007.

Jean-Baptiste MICHEL, Yuan Kui SHEN, Aviva Presser AIDEN, Adrian VERES, Matthew K GRAY, Joseph P PICKETT, Dale HOIBERG, Dan CLANCY, Peter NORVIG, Jon ORWANT *et al.* : Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182, 2011.

Charles MULLER : *Principes et méthodes de statistique lexicale*. Numéro 2. 1980.

Kavi Narayana MURTHY et G Bharadwaja KUMAR : Language identification from small text samples. *Journal of Quantitative Linguistics*, 13(01):57–80, 2006.

Makoto NAGAO et Shinsuke MORI : A new method of n-gram statistics for large number of n and automatic extraction of words and phrases from large text data of japanese. *In Proceedings of the 15th conference on Computational linguistics-Volume 1*, pages 611–615. Association for Computational Linguistics, 1994.

Vlad NICULAE, Marcos ZAMPIERI, Liviu P DINU et Alina Maria CIOBANU : Temporal text ranking and automatic dating of texts. *In EACL*, pages 17–21, 2014.

Thomas R NIESLER et Philip C WOODLAND : A variable-length category-based n-gram language model. *In Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, volume 1, pages 164–167. IEEE, 1996.

Alexandros NTOULAS, Marc NAJORK, Mark MANASSE et Dennis FETTERLY : Detecting spam web pages through content analysis. *In Proceedings of the 15th international conference on World Wide Web*, pages 83–92. ACM, 2006.

Brendan O'CONNOR, Michel KRIEGER et David AHN : Tweetmotif : Exploratory search and topic summarization for twitter. *In ICWSM*, pages 384–385, 2010.

Alexandra OLTEANU, Carlos CASTILLO, Nicholas DIAKOPOULOS et Karl ABERER : Comparing events coverage in online news and social media : The case of climate change. *In Proceedings of the Ninth International AAAI Conference on Web and Social Media*, numéro EPFL-CONF-211214, 2015.

Mark PAGEL : Human language as a culturally transmitted replicator. *Nature reviews. Genetics*, 10(6):405, 2009.

Kishore PAPINENI, Salim ROUKOS, Todd WARD et Wei-Jing ZHU : Bleu : a method for automatic evaluation of machine translation. *In Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

Adam PAULS et Dan KLEIN : Faster and smaller n-gram language models. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies-Volume 1*, pages 258–267. Association for Computational Linguistics, 2011.

Eitan Adam PECHENICK, Christopher M DANFORTH et Peter Sheridan DODDS : Characterizing the google books corpus : Strong limits to inferences of socio-cultural and linguistic evolution. *PloS one*, 10(10):e0137041, 2015a.

Eitan Adam PECHENICK, Christopher M. DANFORTH et Peter Sheridan DODDS : Is language evolution grinding to a halt ? : Exploring the life and death of words in english fiction. *CoRR*, abs/1503.03512, 2015b.

Jian PENG, Kim-Kwang Raymond CHOO et Helen ASHMAN : Bit-level n-gram based forensic authorship analysis on social media : Identifying individuals from linguistic profiles. *Journal of Network and Computer Applications*, 70:171–182, 2016.

Alexander M PETERSEN, Joel N TENENBAUM, Shlomo HAVLIN, H Eugene STANLEY et Matjaž PERC : Languages cool as they expand : Allometric scaling and the decreasing need for new words. *Scientific reports*, 2:943, 2012.

Steven T PIANTADOSI : Zipf's word frequency law in natural language : A critical review and future directions. *Psychonomic bulletin & review*, 21(5):1112–1130, 2014.

Octavian POPESCU et Carlo STRAPPARAVA : Semeval 2015, task 7 : Diachronic text evaluation. *In SemEval@ NAACL-HLT*, pages 870–878, 2015.

Sophie PRÉVOST : Diachronie du français et linguistique de corpus : une approche quantitative renouvelée. *Langages*, (1):23–45, 2015.

Anand RAJARAMAN et Jeffrey David ULLMAN : *Mining of Massive Datasets :*. Cambridge University Press, Cambridge, 10 2011. ISBN 9781139058452. URL https://www.cambridge.org/core/books/mining-of-massive-datasets/A06D57FC616AE3FD10007D89E73F8B92.

Carlos RAMISCH, Aline VILLAVICENCIO et Christian BOITET : Multiword expressions in the wild ? : the mwetoolkit comes in handy. *In Proceedings of the 23rd International Conference on Computational Linguistics : Demonstrations*, pages 57–60. Association for Computational Linguistics, 2010.

Jean RAVENEAU : Tufte, edward r.(1990) envisioning information. cheshire, conn., graphic press, 126 p.(isbn 0-961-3921-1-8). *Cahiers de géographie du Québec*, 37(101):395–397, 1993.

Radim REHUREK et Milan KOLKUS : Language identification on the web : Extending the dictionary method. *In CICLing*, pages 357–368. Springer, 2009.

Lakhdar REMAKI et Jean Guy MEUNIER : Un modèle hmm pour la détection des mots composés dans un corpus textuel. *In Actes de la Conférence JADT-2000*, 2000.

Robert J. RIGGS et S. Jack HU : Disassembly liaison graphs inspired by word clouds. *Procedia CIRP*, 7:521 – 526, 2013.

Yannick ROCHAT, Maud EHRMANN, Vincent BUNTINX, Cyril BORNET et Frédéric KAPLAN : Navigating through 200 years of historical newspapers. *In iPRES 2016*, numéro EPFL-CONF-218707, 2016.

Ivan A. SAG, Timothy BALDWIN, Francis BOND, Ann COPESTAKE et Dan FLICKINGER : *Multiword Expressions : A Pain in the Neck for NLP*, pages 1–15. Springer Berlin Heidelberg, Berlin, Heidelberg, 2002. ISBN 978-3-540-45715-2. URL http://dx.doi.org/10.1007/3-540-45715-1_1.

James M. SAKODA : A generalized index of dissimilarity. *Demography*, 18(2):245–250, 1981. ISSN 1533-7790.

Bahar SALEHI, Nitika MATHUR, Paul COOK et Timothy BALDWIN : The impact of multiword expression compositionality on machine translation evaluation. *In MWE@ NAACL-HLT*, pages 54–59, 2015.

Gerard SALTON et Michael J MCGILL : Introduction to modern information retrieval. 1986.

Ferdinand de SAUSSURE : Cours de linguistique générale, ed. *C. Bally and A. Sechehaye, with the collaboration of A. Riedlinger, Lausanne and Paris : Payot*, 1916.

Helmut SCHMID : Treetagger| a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, 43:28, 1995.

C. SEIFERT, B. KUMP, W. KIENREICH, G. GRANITZER et M. GRANITZER : On the beauty and usability of tag clouds. *In 2008 12th International Conference Information Visualisation*, pages 17–25, July 2008.

Christin SEIFERT, Eva ULBRICH et Michael GRANITZER : *Word Clouds for Efficient Document Labeling*, pages 292–306. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

C. SHANNON : A mathematical theory of communication. *Bell system technical journal*, 27, 1948.

Grigori SIDOROV, Francisco VELASQUEZ, Efstathios STAMATATOS, Alexander GELBUKH et Liliana CHANONA-HERNÁNDEZ : Syntactic dependency-based n-grams : More evidence of usefulness in classification. *In International Conference on Intelligent Text Processing and Computational Linguistics*, pages 13–24. Springer, 2013.

Christian SIEFKES, Fidelis ASSIS, Shalendra CHHABRA et William S YERAZUNIS : Combining winnow and orthogonal sparse bigrams for incremental spam filtering. *In PKDD*, volume 4, pages 410–421. Springer, 2004.

James SINCLAIR et Michael CARDEW-HALL : The folksonomy tag cloud : when is it useful ? *Journal of Information Science*, 34(1):15–29, 2008.

Amit SINGHAL : Modern information retrieval : A brief overview. *IEEE Data Eng. Bull.*, 24 (4):35–43, 2001.

Manhung SIU et Mari OSTENDORF : Variable n-grams and extensions for conversational speech language modeling. *IEEE Transactions on Speech and Audio Processing*, 8(1):63–75, 2000.

Tristan SNOWSILL, Ilias FLAOUNAS, Tijl DE BIE et Nello CRISTIANINI : Detecting events in a million new york times articles. *Machine Learning and Knowledge Discovery in Databases*, pages 615–618, 2010a.

Tristan SNOWSILL, Florent NICART, Marco STEFANI, Tijl DE BIE et Nello CRISTIANINI : Finding surprising patterns in textual data streams. *In Cognitive Information Processing (CIP), 2010 2nd International Workshop on*, pages 405–410. IEEE, 2010b.

Dae-Neung SOHN, Jung-Tae LEE et Hae-Chang RIM : The contribution of stylistic information to content-based mobile spam filtering. *In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 321–324. Association for Computational Linguistics, 2009.

Efstathios STAMATATOS : A survey of modern authorship attribution methods. *Journal of the Association for Information Science and Technology*, 60(3):538–556, 2009.

Efstathios STAMATATOS *et al.* : Ensemble-based author identification using character n-grams. *In Proceedings of the 3rd International Workshop on Text-based Information Retrieval*, pages 41–46, 2006.

Efstathios STAMATATOS, Nikos FAKOTAKIS et George KOKKINAKIS : Automatic text categorization in terms of genre and author. *Computational linguistics*, 26(4):471–495, 2000.

Luc STEELS : Modeling the cultural evolution of language. *Physics of Life Reviews*, 8(4):339–356, 2011.

Luc STEELS : Do languages evolve or merely change ? *Journal of Neurolinguistics*, 2016. ISSN 0911-6044. URL http://www.sciencedirect.com/science/article/pii/S091160441630077X.

Luc STEELS et Frédéric KAPLAN : Spontaneous lexicon change. *In Proceedings of the 17th International Conference on Computational Linguistics - Volume 2*, COLING '98, pages 1243–1250, Stroudsburg, PA, USA, 1998. Association for Computational Linguistics. URL http://dx.doi.org/10.3115/980432.980772.

Hidayet TAKÇI et Tunga GÜNGÖR : A high performance centroid-based classification approach for language identification. *Pattern Recognition Letters*, 33(16):2077–2084, 2012.

Andrija TOMOVIĆ et Predrag JANIČIĆ : A variant of n-gram based language classification. *AI* IA 2007 : Artificial Intelligence and Human-Oriented Computing*, pages 410–421, 2007.

Andrija TOMOVIĆ, Predrag JANIČIĆ et Vlado KEŠELJ : n-gram-based classification and unsupervised hierarchical clustering of genome sequences. *Computer methods and programs in biomedicine*, 81(2):137–153, 2006.

Xiang TONG et David A EVANS : A statistical approach to automatic ocr error correction in context. *In Proceedings of the fourth workshop on very large corpora*, pages 88–100, 1996.

Pedro A TORRES-CARRASQUILLO, Douglas A REYNOLDS et John R DELLER : Language identification using gaussian mixture model tokenization. *In Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 1, pages I–757. IEEE, 2002.

Erik TROMP et Mykola PECHENIZKIY : Graph-based n-gram language identification on short texts. *In Proc. 20th Machine Learning conference of Belgium and The Netherlands*, pages 27–34, 2011.

Yulia TSVETKOV et Shuly WINTNER : Extraction of multi-word expressions from small parallel corpora. *Natural Language Engineering*, 18(4):549–573, 2012.

Matthew TURK et Alex PENTLAND : Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991a.

Matthew A TURK et Alex P PENTLAND : Face recognition using eigenfaces. *In Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on*, pages 586–591. IEEE, 1991b.

Fernanda B. VIÉGAS et Martin WATTENBERG : Timelines : Tag clouds and the case for vernacular visualization. *interactions*, 15(4):49–52, juillet 2008.

Zeev VOLKOVICH, Valery KIRZHNER, Alexander BOLSHOY, Eviatar NEVO et Abraham KOROL : The method of n-grams in large-scale clustering of dna texts. *Pattern recognition*, 38(11): 1902–1912, 2005.

R. VUILLEMOT, T. CLEMENT, C. PLAISANT et A. KUMAR : What's being said near "martha" ? exploring name entities in literary text collections. *In 2009 IEEE Symposium on Visual Analytics Science and Technology*, pages 107–114, Oct 2009.

Fredrik WAHLBERG, Lasse MÅRTENSSON et Anders BRUN : Large scale continuous dating of medieval scribes using a combined image and language model. *In Document Analysis Systems (DAS), 2016 12th IAPR Workshop on*, pages 48–53. IEEE, 2016.

Chi WANG, Marina DANILEVSKY, Nihit DESAI, Yinan ZHANG, Phuong NGUYEN, Thrivikrama TAULA et Jiawei HAN : A phrase mining framework for recursive construction of a topical hierarchy. *In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 437–445. ACM, 2013.

William S-Y WANG, Jinyun KE et James W MINETT : Computational studies of language evolution. *Computational linguistics and beyond*, pages 65–106, 2004.

Xuerui WANG, Andrew MCCALLUM et Xing WEI : Topical n-grams : Phrase and topic discovery, with an application to information retrieval. *In Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 697–702. IEEE, 2007.

Z WEI, JH CHAUCHAT et D MIAO : Comparing different text representation and feature selection methods on chinese text classification using character n-grams. *Journées Internationnales d'Analyse des Données Textuelles*, pages 1175–1186, 2008.

Gerhard WEIKUM, Johannes HOFFART, Ndapandula NAKASHOLE, Marc SPANIOL, Fabian M SUCHANEK et Mohamed Amir YOSEF : Big data methods for computational linguistics. *IEEE Data Eng. Bull.*, 35(3):46–64, 2012.

Cui WEIWEI, Wu YINGCAI, Liu SHIXIA, Michelle Zhou FURU, Wei and et Qu HUAMIN : Context-preserving, dynamic word cloud visualization. *IEEE Computer Graphics and Applications*, 30(6):42–53, 2010.

Ingrid WESTIN et Christer GEISLER : A multi-dimensional study of diachronic variation in british newspaper editorials. *International Computer Archive of Modern and Medieval English*, (26):133–152, 2002.

Derry Tanti WIJAYA et Reyyan YENITERZI : Understanding semantic change of words over centuries. *In Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversiTy on the social web*, pages 35–40. ACM, 2011.

Yingcai WU, Thomas PROVAN, Furu WEI, Shixia LIU et Kwan-Liu MA : Semantic-preserving word clouds by seam carving. *Computer Graphics Forum*, 30(3):741–750, 2011.

Ruifeng XU, Jiyun ZHOU, Bin LIU, Yulan HE, Quan ZOU, Xiaolong WANG et Kuo-Chen CHOU : Identification of dna-binding proteins by incorporating evolutionary information into pseudo amino acid composition via the top-n-gram approach. *Journal of Biomolecular Structure and Dynamics*, 33(8):1720–1730, 2015.

Guangbing YANG, Dunwei WEN, Nian-Shing CHEN, Erkki SUTINEN *et al.* : A novel contextual topic model for multi-document summarization. *Expert Systems with Applications*, 42 (3):1340–1352, 2015.

Chun YUAN, Ni LAO, Ji-Rong WEN, Jiwei LI, Zheng ZHANG, Yi-Min WANG et Wei-Ying MA : Automated known problem diagnosis with event traces. *In ACM SIGOPS Operating Systems Review*, volume 40, pages 375–388. ACM, 2006.

Marcos ZAMPIERI, Alina Maria CIOBANU, Vlad NICULAE et Liviu P DINU : Ambra : A ranking approach to temporal text classification. *In SemEval@ NAACL-HLT*, pages 851–855, 2015.

Richard ZENS et Hermann NEY : N-gram posterior probabilities for statistical machine translation. *In Proceedings of the Workshop on Statistical Machine Translation*, pages 72–77. Association for Computational Linguistics, 2006.

George ZIPF : *The Psychobiology of Language : An Introduction to Dynamic Philology*. M.I.T. Press, Cambridge, Mass., 1935.

Marc A ZISSMAN et Elliot SINGER : Automatic language identification of telephone speech messages using phoneme recognition and n-gram modeling. *In Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, volume 1, pages I–305. IEEE, 1994.

Vincent BUNTINX

Avenue Victor-Ruffy 59, 1012 Lausanne
Né le 2 décembre 1983 à Bruxelles
Belge, Célibataire

Natel: +41 (0)78 691 70 10
Mail: vbuntinx@shogaku.ch
Site: http://www.shogaku.ch

Objectif

Je suis passionné par les domaines des humanités digitales, des mathématiques, de l'informatique et de la physique. J'apprécie le travail en équipe et les défis complexes me stimulent, particulièrement lorsqu'ils font appel à une approche transdisciplinaire.

Compétences professionnelles

Aptitudes professionnelles

- Data mining, text mining et linguistique computationnelle.
- Programmation informatique, data visualization et web scrapping.
- Recherche opérationnelle, optimisation, statistique et probabilités.
- Big data, analyse des réseaux et théorie des graphes.
- Modélisation actuarielle, analyse et gestion des risques.
- Ingénierie de la décision et aide multicritère à la décision.
- Gestion de projets, logistique et analyse de la demande.
- Gestion d'équipe dans un contexte multiculturel.

Aptitudes générales

- Adaptation, créativité, autonomie, responsabilité, sociabilité, organisation, persévérance.

Expériences professionnelles

Collaborateur scientifique / Assistant-doctorant à l'EPFL — *2013 - 2017*

- Recherche en humanités digitales, analyse de données textuelles, développement d'outils en Python et PHP/JavaScript, conduite et gestion des projets étudiants en informatique.

Expérience multidisciplinaire — *2012 - 2013*

- Bioinformaticien à l'Institut Suisse de Bioinformatique (4 mois).
- Analyste programmeur à l'Université de Lausanne (6 mois).

Actuaire ASA chez Figeas SA — *2008 - 2012*

- Analyse des risques, tarification des produits, calcul des provisions techniques et élaboration des modèles actuariels afférents au Swiss Solvency Test (évaluation des scénarios, projection du bilan, convolution des risques, évaluation de produits financiers dérivés par la méthode de Monte Carlo et de l'arbre binomial, modélisation des grands sinistres).

Formation

Doctorat en Humanités Digitales — *2014 - 2017*

- Lieu : EPFL - École Polytechnique Fédérale de Lausanne.
- Programme doctoral : EDMT – École Doctorale du Management de la Technologie.
- Thèse : "Analyse multi-échelle de n-grammes sur 200 années d'archives de presse".
- Directeur de thèse : Frédéric Kaplan (EPFL) et Aris Xanthos (UNIL - Université de Lausanne).

Master en Sciences Actuarielles — *2006 - 2008*

- Lieu : ULB - Université Libre de Bruxelles.
- Grade : Distinction.
- Mémoire : "Constitution et évaluation d'un portefeuille d'options dont la gestion est basée sur l'aide multicritère à la décision".
- Directeurs de mémoire : Bertrand Mareschal (ULB) et Yves De Smet (ULB).

Licence en Sciences Physiques (équivalence Master) — *2002 - 2006*

- Lieu : ULB - Université Libre de Bruxelles.
- Grade : Distinction.
- Mémoire : "Étude moléculaire du transfert d'électron lors de la collision H et He+".
- Directeur de mémoire : Nathalie Vaeck (ULB).

Divers

Informatique

- Logiciels : Microsoft Office, LibreOfffice, OpenOffice, MathLab, Mathematica, Latex, Decision Lab.
- Langages de programmation : Python, R, SQL, HTML, CSS, PHP, JavaScript, Visual Basic.

Langues

- Français (langue maternelle).
- Anglais (niveau intermédiaire).

Loisirs

- Karaté - 3ème Dan (enseignant au club Shogaku).
- Iaido et Kenjutsu - 1er Dan (sabre japonais).
- Yoga et méditation.

Associations

- Membre de la WJKA (World JKA Karate Alliance).
- FEI (Fédération Européenne de Iai).

Références

- Frédéric Kaplan, directeur du laboratoire des humanités digitales à l'EPFL.
- Jean Cochet, actuaire responsable chez Figeas SA.
- Patrick Staeger, enseignant et ancien formateur d'enseignants.