Analyse de données textuelles et sciences sociales : application et comparaison de deux outils, Calliope et Alceste

Sandrine Clérisse

CONSERVATOIRE NATIONAL DES ARTS ET METIERS

Ecole Management et Société-Département CITS

INTD

MEMOIRE pour obtenir le Titre enregistré au RNCP
"Chef de projet en ingénierie documentaire"
Niveau I

Présenté et soutenu par
Sandrine Clérisse

le 25 novembre 2015

Analyse de données textuelles et sciences sociales

Application et comparaison de deux outils, Calliope et Alceste

Jury : Maryse Carmes (CNAM), Mathilde de Saint-Léger (CNRS)

Promotion 45

« Hâtez-vous lentement, et sans perdre courage,
Vingt fois sur le métier remettez votre ouvrage,
Polissez-le sans cesse, et le repolissez,
Ajoutez quelquefois, et souvent effacez. »

Nicolas Boileau, *L'Art poétique*, Chant I

« Il n'y a de vision *que* perspective,
il n'y a de "connaissance" *que* perspective »

Friedrich Nietzsche, *Généalogie de la morale*

Remerciements

Je tiens à remercier les personnes qui ont rendu possible et suivi ce travail : Maryse Carmes, pour son soutien sans faille, ses questions pertinentes et stimulantes ainsi que ses multiples références ; Mathilde de Saint-Léger pour son attention critique et son encadrement discret mais efficace – les maladresses présentes dans ce mémoire demeurent de ma responsabilité, tant il est parfois difficile et long d'assimiler un conseil… –, Sarah Gensburger, pour son enthousiasme et sa détermination inébranlables et qui, la première, m'a proposé de collaborer à son projet de recherche et enfin, Sophie Duchesne, pour avoir pris le temps de m'expliquer (et de me réexpliquer…) la méthodologie d'Alceste.

Notice

CLÉRISSE, Sandrine. Analyse de données textuelles et sciences sociales. Application et comparaison de deux outils, Calliope et Alceste – Mémoire pour le titre professionnel de niveau I "Chef de projet en ingénierie documentaire", Paris : INTD-CNAM, 2015, 233 p., promotion 45

L'accès à des corpus textuels, d'une taille inhabituelle et d'un type nouveau, se voit largement facilité de nos jours (politique d'ouverture des données publiques, partage des données avec le web sémantique, sources web de contenus tels que les réseaux sociaux, etc.). Cette plus grande commodité d'accès nécessite le recours à des méthodes et des outils informatisés capables de traiter et d'analyser cette masse d'informations. L'analyse de données textuelles fait partie de ces méthodes d'exploration des données, basées sur des principes statistiques et linguistiques. La lecture à distance des données qu'elle permet, tout en étant fondée sur elles, représente une opportunité pour la démarche scientifique : les pistes interprétatives que l'analyse lexicométrique propose sont autant de nouveaux questionnements scientifiques. Pourtant, ces méthodes sont encore peu pratiquées en sciences sociales. Notre travail vise à montrer l'intérêt que peut représenter l'intégration d'une telle démarche au sein d'un projet de recherche en sociologie, en procédant à l'analyse lexicométrique du vocabulaire du phénomène mémoriel. Pour cela, nous avons appliqué deux logiciels d'analyse de données textuelles, Calliope et Alceste, au même corpus de données : les déclarations d'associations au *Journal officiel*. Ce travail se veut doublement exploratoire : établir une méthodologie d'analyse des données qui en révèle les risques potentiels pour mieux les déjouer et proposer des pistes de réflexion dans le cadre d'un prolongement du projet de recherche. L'un des apports de ce travail est de montrer que le recours à de multiples outils d'analyse de données textuelles constitue une réelle complémentarité, bénéfique aux projets de recherche, en sciences sociales notamment.

Descripteurs : Analyse de données ; Analyse de mots associés ; Analyse lexicale ; Analyse statistique ; Analyse multifactorielle ; Classification hiérarchique ; Traitement des données ; Lemmatisation ; Analyse comparative ; Méthodologie ; Humanités numériques ; Sciences sociales ; Sciences humaines ; Recherche scientifique ; Document textuel ; Cluster ; Cooccurrence ; Donnée

---

The access to textual corpora, of unusual size and new type, is largely being facilitated nowadays (policies for opening public data, data sharing with the semantic Web, Web sources of contents such as social networks, etc). This greater accessibility requires the use of methods and computerized tools able to treat and analyze this mass of information. Textual data analysis is one of these data exploration methods, based on statistical and linguistic principles. The remote reading of data which it allows, although founded on them, represents an opportunity for the scientific approach: the interpretative tracks that it provides opens to new questionings. However these methods are still hardly used in social sciences. Our work aims at showing the interest which the integration of such an approach within a sociological research project can represent, while carrying out a lexicometric analysis of the vocabulary of the memory phenomenon. For that, we applied two textual data analysis softwares, Calliope and Alceste, to the same corpus of data: declarations of associations in the French *Journal officiel*. This work wants to be exploratory in two ways: to establish a methodology of data analysis which reveals the possible hazards, with a view to avoid them, and to offer thoughtful directions to extend the research project. One of the contributions of this work is to show that the use of multiple tools for analysing textual data

constitutes a real complementarity, useful in the context of research projects, in the social sciences in particular.

Keywords : Data Analysis ; Associated-Word Method ; Lexical Analysis ; Statistical Analysis ; Multifactorial Analysis ; Hierarchical Clustering ; Data Processing ; Lemmatization ; Comparative Analysis ; Methodology ; Digital Humanities ; Social Sciences ; Humanities ; Scientific Research ; Textual Record ; Cluster ; Cooccurrence ; Data

Table des matières

Liste des figures

Introduction

« Les logiciels permettant le traitement de gros corpus textuels sont encore peu utilisés en France par les sociologues. » Ce constat, énoncé par des sociologues et par lequel débute l'ouvrage collectif *Analyse textuelle en sociologi*e [31, DEMAZIERE et al.], pourrait constituer le point de départ à cette étude. Ce constat fourmille de paradoxes.
Le premier se révèle en mettant en regard, d'un côté, des recherches fortement préoccupées par le recueil de matériaux textuels, comme support à l'analyse (entretiens, questions ouvertes, données collectées, notamment textuelles, en situation d'observation, etc.) et, de l'autre, une distance à l'égard de l'analyse statistique de données textuelles. Ainsi que le soulignent Demazière *et al.*, « d'un côté, les matériaux non numériques, et notamment langagiers, dominent très largement dans la production sociologique, et de l'autre les exploitations de ces matériaux s'appuient dans la grande majorité des cas sur des méthodes manuelles, qui ne peuvent être mises en œuvre que sur des corpus limités en taille, même pour les plus formalisées et codifiées d'entre elles. »

L'une des raisons de cette frilosité des sciences sociales envers les méthodes d' analyse de données textuelles pourrait se situer du côté de l'opposition, sommaire, entre démarche qualitative et démarche quantitative. A la froideur du chiffre, manquant de chair et de discernement, répondrait une analyse en profondeur de la complexité composant la matière sociale. Si la quantification et l'étude des longues séries, notamment en histoire, ont connu un fort engouement au XX^{e} siècle, avec une apogée durant les années 60 et 70, se substitue, à partir de la fin des années 70, une période de désenchantement et de méfiance. Les excès de la démarche quantitative informatisée, sa prétention à capter « objectivement » les objets sociaux suscitent de nombreuses critiques. « Excès [notent C. Lemercier et C. Zalc] "avec la croyance qu'il suffisait de rassembler des données par brouettées pour écrire l'histoire et que les conclusions allaient sortir toutes armées de l'ordinateur" » et excès dans l'emploi « d'outils surdimensionnés pour démontrer l'évidence. […] Le temps passé et la sueur versée à saisir, compter et classer valent-ils la peine si le processus aboutit à exhiber des résultats sinon prévisibles, tout au moins attendus ? [83, LEMERCIER et ZALC] »

Ce rejet, ou du moins cette crainte vis-à-vis de méthodes quantitatives « non orthodoxes » (nous entendons par là les méthodes qui ne relèvent pas de la statistique descriptive), s'avère d'autant plus paradoxale que depuis les années 60, les prodigieux progrès des ordinateurs en termes de puissance computationnelle, conjugués à la propagation toujours plus importante des ordinateurs personnels et de leurs usages, permettent de traiter de plus en plus de données, en moins en moins de temps.
Le déploiement des méthodes statistiques d'analyse lexicale s'inscrit dans la longue histoire de ce que l'on nomme « humanités numériques », ce domaine de rencontres multiples entre informatique, technologies de l'information et questions de recherche en sciences humaines et sociales. L'analyse de données textuelles est en effet issue de recherches qui remontent aux années 50 et voit, ces dernières années, ses méthodes et outils connaître un regain d'intérêt, en relation à l'accès facilité à de grands corpus textuels. L'intérêt dont ces méthodes et pratiques font l'objet, tant au niveau académique que politique, n'est que le reflet d'un changement plus profond et plus large, celui de l'entrée de la société dans l'ère numérique. A la fois témoins et acteurs des modifications qui affectent les relations sociales (émergence des réseaux sociaux sur le web notamment), l'organisation économique ou encore les principes de gouvernance, les humanités numériques représentent, à l'échelle académique, un changement majeur dans les modes d'élaboration des objets de recherche, les méthodes de recherche, et plus concrètement, les pratiques scientifiques. Mais ancrées dans une filiation davantage « humaine » que « sociale », selon les termes de M. Dacos [13, DACOS], qui se poursuit de nos jours, les humanités numériques n'auraient ainsi pas encore su convaincre les sociologues de l'intérêt de leur contribution aux projets scientifiques.

Notre travail s'inscrit précisément à l'intersection des sciences sociales et des méthodes et outils statistiques appliqués aux données textuelles. Il n'entend pas pour autant dresser un portrait exclusivement élogieux, et donc biaisé, d'une telle démarche. La question essentielle

qui a initié et guidé ce travail, tout au long de ses nombreuses étapes, est ainsi celle de l'intérêt potentiel qu'offrent de telles pratiques à un projet de recherche en sciences sociales. Nous nous sommes ainsi demandé si l'analyse de données textuelles ne pouvait pas constituer une source d'opportunités pour les sciences sociales, si elle ne représentait pas un des points d'entrée possibles aux faits sociaux. L'élargissement des corpus, que rend possible leur traitement par des outils lexicométriques, ne signifie-t-il pas une prise en compte plus large des faits langagiers, et partant, des faits sociaux ? Autrement dit, en quoi l'alliance de méthodes de statistique lexicale et de questions sociologiques peut-elle constituer un avantage pour la recherche ?
Afin de répondre à ces questions, nous avons émis l'hypothèse, appuyée sur plusieurs lectures [voir notamment 27, BEAUDOIN ; 36, GUERIN-PACE ; 50, MAYAFFRE et 59, REINERT], que l'aide que pouvait représenter une démarche d'analyse statistique textuelle ne saurait être qu'exploratoire. Ni descriptive d'objets sociaux, ni nouveau moyen d'administration de la preuve, ni même explicative de phénomènes, la démarche lexicométrique « ne peut que » proposer des éléments, éventuellement étonnants, parfois concordants, à l'analyse et les soumettre à l'interprétation de l'expert. Si la « lecture de surface » qu'elle opère des innombrables matériaux textuels s'oppose en ce sens à la lecture intensive de l'expert, ce n'est que pour mieux la compléter. Comme le remarquent C. Lemercier et C. Zalc, « compter pour mieux chercher », pour mieux questionner.

Mais cette aide ne saurait être instantanée et l'objectivité dont se parent les méthodes quantitatives, ou plutôt dont certains les dotent, n'est qu'apparente. Une aide, certes, mais une aide construite, élaborée, réfléchie L'analyse lexicométrique ne se résume donc pas à l'usage d'outils, qui eux-mêmes ne se réduisent pas à de simples « presse-boutons ». Tout l'enjeu de notre travail se situe justement dans cette perspective. Pour prolonger la question posée plus haut, il ne s'agit plus tant de savoir si l'analyse de données textuelles peut être utile aux sciences sociales que, plus précisément, à quelles conditions cela est possible. Ce qui explique pourquoi le travail que nous avons mené est profondément exploratoire, et ce, à double titre : tout d'abord, parce que l'application d'une démarche d'analyse de données textuelles est orientée par une finalité exploratoire, comme nous l'avons déjà énoncé, mais également en raison du cadre même du projet auquel nous participons. Non seulement, ce projet de recherche en est à ses premières étapes, mais notre travail lui-même s'inscrit dans un projet de réorientation professionnelle, finalisé par un stage. Le but de ce stage était donc de toucher du doigt (au sens propre comme au figuré, concernant les *Digital Humanities*[1] !) les difficultés qu'implique le traitement de données, notamment textuelles, d'éprouver la nécessité de réfléchir à une méthodologie et d'acquérir les compétences indispensables à la poursuite de notre carrière professionnelle. Il en résulte que les trois mois de stage ont été mis à profit pour tenter de déterminer une méthode adaptée à l'analyse de données textuelles, comprenant notamment les phases, délicates, d'élaboration du corpus. Nécessairement circonscrit dans le temps, ce travail ne prétend être ni exhaustif ni définitif. Il ne constitue qu'une étape, qu'une facette d'un projet plus large et de plus longue haleine.

De quel projet de recherche s'agit-il ?
Le projet de recherche en sciences sociales auquel nous collaborons s'interroge sur l'émergence du phénomène mémoriel en France et sa constitution en objet social. La finalité des travaux de recherche de S. Gensburger, sociologue et porteuse du projet, concerne les usages politiques du passé, et plus précisément, de la question mémorielle. Elle entend montrer que l'analyse de cette question d'un point de vue sociologique est à même d'appréhender les mécanismes et dynamismes sociaux qui y sont à l'œuvre. Cet ancrage sociologique ne dispense pas pour autant d'une analyse du contexte historique dans lequel se situent ces événements et phénomènes.
Ce projet remet en perspective, par leur questionnement, les conceptions couramment admises sur le « phénomène mémoriel ». Celles-ci lui attribuent en effet une origine

[1] Pour une approche des humanités numériques comme proprement digitales, voir [24, URBANIAK] et [12, CLIVAZ].

« communautariste », provenant de la fragmentation dont serait affectée la société française depuis les années 80, et qui se traduirait par une inflation des « revendications mémorielles particularisées ». Selon S. Gensburger, cette lecture méconnaît à la fois le rôle actif joué par l'Etat dans le processus d'institutionnalisation de la mémoire, dans son élaboration en objet social, ainsi que la chronologie dans laquelle s'inscrivent ces événements. « Une telle approche devrait conduire à reprendre à nouveaux frais des notions qui ont aujourd'hui valeur d'évidence comme celles d'"inflation mémorielle", de "concurrence des mémoires", de "devoir de mémoire", de "réconciliation", de "mémoire partagée" ou encore d'"usages politiques du passé". »
Cette attention portée au langage et aux expressions employées pour désigner le phénomène mémoriel n'est pas anodine. Si le langage est un matériau éminemment social, l'étude sociologique des faits langagiers qui se rapportent à la mémoire peut alors contribuer à son éclairage. D'où la mise en place, dans le cadre de ce projet, d'une démarche incluant des outils de traitement et d'analyse du lexique.

De quels matériaux textuels parle-t-on ?
Afin de montrer que la demande sociale suit davantage qu'elle ne prescrit la question mémorielle et son calendrier, le projet se fonde sur l'analyse des déclarations d'associations publiées au *Journal officiel* (JO), de 1945 à nos jours. La masse de données disponibles est telle qu'elle écarte d'emblée toute possibilité d'analyse purement manuelle.
L'étude des déclarations au JO, au moyen de l'analyse de données textuelles, devrait permettre de contraster l'usage des éléments langagiers propres aux associations et tenter de voir comment se distribue(ent) le(s) lexique(s) mémoriel(s), en termes de thématiques, de type d'associations, de temporalité ou encore de répartition géographique. Peut-on voir se dessiner une « inflation » du registre mémoriel ? Quels univers thématiques ou types de discours « disent » la mémoire / les mémoires ? Ces lexiques évoluent-ils dans le temps, dans l'espace ? De quelle manière ?
Il s'agit, plus précisément, de voir si le positionnement des deux pôles du phénomène mémoriel identifiés par l'opinion commune (une demande sociale prescriptrice d'un côté et, de l'autre, des administrations centrales qui suivent) ne s'inverse pas : l'Etat, au travers de ses liens forts aux associations d'anciens combattants, jouerait un rôle de premier plan dans la constitution de la mémoire comme objet social, tandis que les associations « communautaristes » n'en seraient que des acteurs plus tardifs. Une attention particulière sera donc portée aux lexiques de ces deux types d'acteurs et à leurs évolutions.

Quels outils appliquer aux matériaux textuels ?
Pour procéder à l'analyse de ces vocabulaires, deux logiciels de statistique lexicale ont été choisis : Calliope et Alceste. Ces deux outils, fondés sur des méthodes statistiques proches mais distinctes, opèrent une classification des matériaux textuels, en fonction de la proximité (et de la répétition de cette proximité) de certains de leurs éléments. De cette façon, il devrait être possible d'extraire de ces données des spécificités lexicales (univers thématiques selon Calliope, types de discours selon Alceste) et de saisir, au moyen de l'interprétation qui en resituera le contexte institutionnel, politique et social notamment, la dynamique sociale qui y est attestée. Le choix de ces logiciels s'explique par le fait qu'ils sont particulièrement adaptés au type de discours propre aux déclarations : descriptifs, formalisés, peu sophistiqués d'un point de vue linguistique et discursif, ces textes conviennent bien aux postulats statistiques et linguistiques implémentés dans les outils retenus.
Mais, comme nous l'avons déjà dit, notre travail étant exploratoire, il ne vise pas tant à fournir des résultats qu'à tenter d'élaborer une méthode adaptée au traitement et à l'analyse des données du JO. Etant donné la diversité des outils disponibles, et des méthodes sur lesquelles ils sont fondés, faire le choix d'un outil signifie faire le choix d'un certain type de résultats produits. C'est pourquoi l'objectif majeur de notre travail a été d'appliquer deux outils au même corpus, afin de comparer leurs méthodes et résultats et voir ce qui les rapprochait ou les différenciait. Nous essaierons de montrer, dans une perspective comparative, que, loin d'être problématique, cette diversité constitue plutôt une réelle complémentarité. Multiplier les points de vue sur un même matériau devrait permettre de mieux en appréhender la richesse, les multiples facettes et, partant, d'offrir d'autres pistes à

l'interprétation et au questionnement scientifique. Pour reprendre Demazière *et al.*, « [user de méthodes logicielles différentes sur un même corpus peut] servir des lectures variées, et indiquer ainsi la voie à des usages raisonnés et fructueux [31, DEMAZIERE et al.] ».
Pour mener à bien notre étude, nous avons procédé par étapes, chacune nécessitant interrogations et allers-retours, exploration de pistes finalement abandonnées au profit d'autres pistes. Nous tenterons de montrer que ce n'est qu'au prix d'une telle réflexivité et itérativité qu'une démarche d'analyse de données textuelles non seulement est possible mais peut s'avérer bénéfique à un projet de recherche.

La présentation de notre travail, au sein de ce mémoire, se décompose de la manière suivante : dans une première partie, nous poserons le cadre, en présentant tout d'abord le projet de recherche sur le phénomène mémoriel, que dirige S. Gensburger et auquel nous participons. Nous indiquerons notamment les hypothèses de recherche sous-jacentes ainsi que les attentes de la chercheuse vis-à-vis de l'analyse statistique de données textuelles, les questions qui nous ont guidées dans le traitement des données du JO. Nous poursuivrons par un rapide tableau des humanités numériques, notamment des enjeux qui les sous-tendent, en insistant davantage sur l'une de ses branches, l'analyse de données textuelles – sollicitée dans ce travail. Nous tenterons de montrer que les méthodes dites « quantitatives » sont de fait sans cesse entremêlées de questionnements et réflexions qualitatifs et que les outils de lexicométrie, bien loin d'être neutres, représentent de véritables « épistémologies embarquées [31, DEMAZIERE et al.] ». Nous terminerons par une présentation des divers outils relevant de ce domaine, et plus particulièrement des deux outils retenus ici : Calliope et Alceste.
La deuxième partie abordera les questions méthodologiques, non négligeables dans ce type de démarche, avec un accent particulier donné aux phases d'élaboration du corpus et de prétraitement des données. L'étape de production de résultats sera également abordée. Cette partie sera l'occasion de mettre en avant les nombreux allers-retours et les incessants questionnements dont le projet a été l'objet, ainsi que les difficultés qui se sont présentées et les précautions à prendre en considération.
Enfin, une dernière partie exposera les résultats obtenus au moyen des deux outils d'analyse textuelle, qui soulignera leur complémentarité, dans une dimension comparative.
Des conclusions seront tirées de ce travail, à l'échelle du projet lui-même – en nous fondant sur une analyse critique de la méthode adoptée et des résultats qui en sont issus – mais aussi à l'échelle individuelle et professionnelle, sur ce que peut être la position d'un ingénieur en analyse de données textuelles au sein d'un projet de recherche.

Nous aimerions conclure cette introduction sur la dimension collective du projet auquel nous collaborons. Ce projet implique – a impliqué et impliquera[2] – en effet plusieurs personnes aux compétences variées. Caractéristique en cela des projets menés au sein des humanités numériques, il témoigne aussi de la difficulté qu'il y a à vouloir traiter et analyser des masses de données importantes, qui plus est, textuelles. La quantité mais également la nature même des données, leur disponibilité (ou absence de disponibilité…), leur structuration (ou absence de structuration…), etc. font des projets « numérico-qualitatifs » de vastes entreprises, nécessitant de multiples ressources, en termes de compétences, de financement et finalement… de temps.

[2] Outre S. Gensburger, porteuse du projet, collaborent à ce projet : M. de Saint-Léger, conceptrice de Calliope et dont l'expertise en analyse de données textuelles est déterminante ; E. Benaïssa, spécialiste du langage Python, qui a procédé, dans le cadre d'un stage, à la structuration et au nettoyage des données des fichiers du JO ; Brian Chauvel, doctorant en sociologie, qui participe notamment au volet prosopographique du projet.

Première partie
Problématique et état de l'art

1 Présentation du projet de recherche

1.1 Analyse du phénomène mémoriel

1.1.1 Une analyse sociologique

Avant de présenter le projet de recherche auquel nous collaborons, nous allons commencer par introduire la notion de mémoire qui y est à l'œuvre.
Le projet de recherche mené par la sociologue Sarah Gensburger, et dans lequel notre étude sur l'analyse des données textuelles s'inscrit, porte sur les usages politiques du passé, et plus précisément, sur la construction de la mémoire en question sociale[3]. Ce projet est à la fois plus ancien et plus large que la partie nous concernant. Plus ancien en ce sens qu'il constitue une poursuite des travaux antérieurs de la chercheuse et plus large, car d'autres aspects, d'autres dimensions du projet sont en cours d'analyse ou feront l'objet de recherches futures (notamment une analyse prosopographique des acteurs du phénomène mémoriel).
Les travaux de Sarah Gensburger relèvent d'une sociologie de la mémoire, qui emprunte à la fois aux pratiques historiennes et aux sciences politiques pour mieux s'en démarquer. Puisant d'un côté dans les manières dont on écrit ou commémore l'histoire collective et de l'autre, dans l'analyse des politiques publiques, ces recherches introduisent, notamment dans le cadre de sa thèse [1, GENSBURGER], une dimension empirique, en étudiant, sur le terrain, les différents acteurs du processus mémoriel.
Placer l'étude du processus mémoriel au cœur des objets sociologiques permet en effet de rendre compte de la complexité des dynamiques sociales qui y sont à l'œuvre que la simple reconnaissance de l'existence, depuis les années 80, d'une l'inflation mémorielle empêche de questionner. « La notion de "devoir de mémoire" est aujourd'hui en France un lieu commun, un poncif de l'évocation du passé dans l'espace public [5, GENSBURGER et LAVABRE] » qui ne suscite pourtant pas de questionnement de la part des médias ni de la communauté des historiens. Or, cette inflation mémorielle – et la critique qu'elle a suscitée, le « droit à l'oubli » – demandent à être interrogées, notamment d'un point de vue sociologique, car ces notions posent une question essentielle : « peut-on agir sur la mémoire ? ».
L'opposition entre une « mémoire obligée » et une « mémoire légitime » relèvent, en fait, d'une même posture qui définit la mémoire en « termes strictement normatifs » sans pour autant expliquer comment « la » mémoire pourrait être manipulée, orientée. Cette posture ne rend ainsi pas compte des différents « niveaux d'expression de la mémoire – privés ou publics, individuels ou collectifs, spontanés ou stratégiques, nés de la volonté des acteurs ou hérités des cadres sociaux. [5, GENSBURGER et LAVABRE] »
Une approche sociologique du phénomène mémoriel permet alors de dépasser l'acceptation d'une action possible sur la mémoire (dans un sens ou dans l'autre, pour la contraindre ou la libérer) en se posant la question du « comment ». Il faut pour cela replacer la mémoire dans ses « cadres sociaux » qui seuls permettent d'appréhender l'individuel et le collectif et les multiples relations qu'ils entretiennent.
C'est pourquoi une sociologie de la mémoire qui veut rendre compte des différentes manifestations empiriques de la mémoire ne doit pas se limiter aux seules manifestations institutionnelles des usages du passé. « La réponse à la question "peut-on agir sur la mémoire" suppose a minima une distinction entre une "mémoire historique" ou usages politiques du passé, une norme mémorielle officielle, visant l'homogénéisation des

[3] Nous nous appuyons, pour la présentation du projet de recherche, sur divers écrits de Sarah Gensburger. Cette recherche étant en cours, certains de ces écrits sont en cours de publication tandis que d'autres ont été rédigés dans le cadre de réponse à des appels à projets. Ce qui explique que nous ne puissions citer précisément toutes les sources. D'autres publications sont néanmoins disponibles (voir Bibliographie).

représentations, d'une part, et une mémoire vive, souvenirs d'un passé vécu ou transmis portés par les individus, de l'autre. [5, GENSBURGER et LAVABRE] »

Si la question mémorielle est importante à analyser, c'est qu'elle implique, de fait, une question plus fondamentale, qui est celle de la « mémoire en démocratie » : « Comment penser une « mémoire démocratique » ou d'une mémoire « en démocratie » (Lavabre, à paraître) ? Faut-il mettre l'accent sur la pluralité des mémoires et l'existence d'une gestion apaisée, c'est-à-dire politique, des conflits d'interprétation du passé ? Et juger alors que la démocratisation de la société passe par l'expression du conflit des interprétations du passé dans l'espace public, la mise en récit publique de toutes les expériences et souvenirs ? Ou faut-il considérer que la mémoire en démocratie passe par le consensus dans les interprétations et les évocations du passé, par une forme de « juste mémoire » que les historiens auraient notamment pour mission de nourrir ? [5, GENSBURGER et LAVABRE] »

Ce faisant, l'approche sociologique de la mémoire a le mérite de rappeler la dimension proprement politique du débat mémoriel. Les termes mêmes des débats contemporains sur la mémoire doivent alors questionnés : Dans quelle mesure peut-on parler de « politique de la mémoire » ? Quels effets ont les commémorations publiques sur la société ? Peut-on établir des liens entre « montée des communautarismes » et « inflations des commémorations » ?
Et « le » vocabulaire mémoriel étudié : De quel vocabulaire s'agit-il ? Que désigne-t-il ? Qui l'emploie ? De quelle manière ? Peut-on constater des spécificités par type d'acteur ? Des spécificités temporelles, géographiques ?

1.1.2 Le phénomène mémoriel en France : Rôle du ministère des Anciens combattants

Depuis le début des années 1980, le terme de mémoire a inspiré un nombre exponentiel de travaux scientifiques dont, de manière exemplaire, la publication à la fin des années 80, de l'ouvrage de Pierre Nora, les *Lieux de mémoire*. Selon Sarah Gensburger, cette approche historienne des politiques de la mémoire se révèle pourtant paradoxalement a-historique. « Jusqu'ici, c'est principalement le regard du chercheur, le plus souvent historien, qui définit les contours de ce qui relève, ou non, de ces dites politiques. Pierre Nora est ainsi allé jusqu'à forger le néologisme de "lieu de mémoiriser" pour décrire l'opération à travers laquelle l'historien constitue le "lieu de mémoire" comme tel (Nora, 1997). Le fait de "lieu de mémoiriser" permet à l'historien de simultanément qualifier et analyser, au final d'instituer, ces politiques dites de la "mémoire" caractéristiques de ce que Pierre Nora nomme l'"ère de la commémoration" (Nora, 1997). Dans ce qui suit, il s'agit de renverser la perspective pour partir d'abord de la qualification de l'action publique par ceux qui la font, l'Etat, ses élus et ses administrations. »
Replacée dans sa dimension historique et sociale, la « politique de la mémoire » se révèle comme un domaine d'activité de l'Etat relativement récent en France. Le terme « mémoire » dans l'organigramme du gouvernement n'a été introduit qu'en 2014 en France (création d'un Secrétariat d'Etat aux Anciens Combattants et à la Mémoire). Ce changement marque en fait « l'aboutissement d'un processus continu d'institutionnalisation d'une "politique de la mémoire" » : en 2008, cinq rapports publics sont ainsi publiés sur les "questions mémorielles". Cet intérêt de l'Etat pour les questions mémorielles est ancien. Dès 1917, un sous-secrétaire d'Etat pour les anciens combattants est créé. « Mais la principale mission du ministère demeure la réparation des souffrances endurées par les soldats et leurs familles. Un fonctionnement partenarial, proche de la cogestion, se met ainsi en place avec ses ressortissants, constitués peu à peu en clientèle (Rémond, 1955). L'activité d'évocation publique du passé, via les commémorations et les monuments, ne vaut alors que parce que ce passé fut directement vécu, et ressenti, par des individus, ex-combattants ou familles civiles, présents en nombre dans la société française. La commémoration est alors considérée comme une forme de réparation symbolique. En conséquence, elle est qualifiée de politique du "souvenir", le terme renvoyant très directement à des individus en chair et en os. »

La Seconde Guerre mondiale ne constitue pas non plus l'émergence d'une politique publique de la mémoire, énoncée comme telle. Ainsi, « en 1945, la gestion des suites de la guerre est confiée principalement au Ministère des Anciens Combattants et Victimes de Guerre (MACVG), qui s'inscrit dans la continuité de son prédécesseur. Il a pour principale mission de "réparer", c'est à dire de satisfaire les besoins des anciens combattants, et de superviser, sous le pilotage direct de la Présidence de la République, la création et l'entretien des cimetières de guerre et de monuments commémoratifs. Son activité se limite à l'érection des lieux et à l'organisation de cérémonies commémoratives. Elle n'est alors aucunement qualifiée d'un terme spécifique. L'écriture de l'histoire de la guerre relève elle d'un service distinct. Elle est confiée au Comité d'histoire de la Seconde Guerre mondiale, placé sous l'autorité directe de la Présidence du conseil et intégré au Secrétariat Général du gouvernement. Ni le MACVG, ni le Comité d'histoire de la Seconde Guerre mondiale ne parle de "mémoire" ni ne prétend en relever. »

L'explicitation d'une politique de la mémoire en tant que telle apparaît, selon Sarah Gensburger dans les années 70, à la suite d'une réduction importante des effectifs du ministère des Anciens combattants, au point de mettre en péril l'existence même du ministère. « Il devient urgent de trouver un nouveau "créneau" comme plusieurs responsables du ministère le formulent eux-mêmes. Le "créneau" envisagé devra pouvoir être, à terme, déconnecté des anciens combattants en chair et en os et pouvoir ainsi leur survivre. » D'où l'émergence de la notion de « mémoire » comme nouveau « créneau » potentiel du ministère des Anciens combattants.

En 1982, le ministère des Anciens Combattants crée la Mission permanente aux commémorations et à l'information historique, qui prend appui sur des commissions départementales et sur les liens du ministère avec le monde associatif. « C'est donc par un appel aux associations que le Ministère développe ce nouveau "créneau" ». Au fil du temps, et de la baisse des anciens combattants, ces organismes évoluent mais conservent, tout en les institutionnalisant, les "actions de mémoire" comme raison d'être. Dernier avatar de cette politique, la Direction de la mémoire, du patrimoine et des archives" (DMPA), créée en 1999, qui a pour mission de « [conduire] des actions dans les domaines de la politique de la mémoire des guerres et des conflits contemporains et de la mise en valeur des lieux de mémoire et des monuments historiques ».

Cette politique de la mémoire, initiée au sein du ministère des Anciens combattants, se diffuse alors dans les diverses administrations centrales. « En effet, si les années 80 sont celle du déploiement de la "mémoire" parmi les missions du MACVG, elles sont également celles de la naissance des "célébrations" au sein du ministère de la Culture. Ce dernier crée un comité aux Célébrations nationales en 1974. Cette activité va à son tour se transformer au fil des années 90 pour désormais s'institutionnaliser autour des termes de "mémoire" et de "commémoration". La "politique de la mémoire" devient l'objet d'une concurrence constante entre administrations. » Ainsi, « davantage qu'une "concurrence des mémoires" », S. Gensburger préfère parler de « concurrence entre administration des Anciens combattants, devenue de la Défense, et administration de la Culture, [concurrence] qui a fait émergé un espace social de la "mémoire", un marché, avec ses ressources et intérêts, pour poursuivre la métaphore économiste. Au fil des années 80, la "mémoire" devient une ressource. A rebours des analyses dominantes, qui omettent d'historiciser la catégorie même de "politique de la mémoire", c'est l'Etat qui avant même le monde associatif et les acteurs sociaux a posé les "questions mémorielles". »

Le développement du phénomène mémoriel, en France, relève donc d'une démarche active de la part des administrations centrales. Et « contrairement aux analyses aujourd'hui dominantes qui décrivent l'Etat comme victime d'une demande sociale débridée et particulariste en matière de "mémoire", [le projet de recherche de S. Gensburger] met en évidence un mécanisme opposé. »

L'étude de la chronologie de la création de journées nationales de commémoration[4] permet également de contredire ce point de vue d'un Etat passif vis-à-vis de « revendications particularistes ». Ainsi, la multiplication contemporaine des journées de commémoration doit être reliée, « non tant à une soi-disant "demande sociale" qu'à la création d'une

[4] Voir en annexe (p. 221) la liste de ces journées nationales de commémoration.

administration dédiée. En effet, l'ancêtre de la DMPA, la Mission permanente aux commémorations et à l'information historique est créé en 1982. Or cette année marque le début de ce qui est aujourd'hui qualifiée d'"inflation des commémorations". De même, c'est seulement quelques mois après la création de la DMPA et, avec elle, l'institutionnalisation définitive d'une action gouvernementale en la matière que le nombre de journées nationales de commémoration connaît une croissance exponentielle, qui n'a pas cessé depuis. Entre 1983 et 2014, 11 journées de commémoration nationale ont été créées par décret ou par la loi contre seulement 3 entre 1914 et 1982. »

1.2 Problématique et hypothèses de recherche

Le projet de recherche sur la « politique mémorielle » entend donc questionner deux croyances particulièrement fortes, largement répandues :
- « Au cours des années 80, la fragmentation de la société française aurait donné lieu à l'éclatement de mémoires particulières (notamment ethniques) qui ont conduit à une demande sociale à l'égard de l'Etat, Etat qui a suivi cette demande en s'engageant sur le terrain deS mémoireS au lieu d'être du côté de l'histoire. C'est ce que Pierre Nora qualifie de "tyrannie de la mémoire".
- Outre qu'elle a changé la nature de la politique de la mémoire, en en consacrant l'éclatement en faveur des particularismes, cette évolution aurait aussi donné lieu à une croissance quantitative : "l'inflation mémorielle", "la poussée mémorielle" etc. selon les auteurs. »
Il s'agit donc de se demander en quoi une double méconnaissance (concernant le rôle de l'Etat et celui de « revendications particularistes ») ne permet pas de comprendre l'institutionnalisation du phénomène mémoriel. « La recherche en cours fait l'hypothèse que la lecture [couramment répandue de ce phénomène] méconnaît à la fois le rôle de l'Etat et passe à côté de la question centrale qui est celle du comment ces éventuelles « revendications particularistes » en sont venues à s'exprimer à travers la "mémoire" et non sur d'autres terrains (ou en complément d'autres terrains) comme par exemple les "discriminations", la reconnaissance des communautés etc. ».

L'hypothèse élaborée par Sarah Gensburger se décline de la façon suivante :

1. **Appréhension quantitative et chronologique erronée du phénomène mémoriel**
Sur un plan quantitatif d'abord, l'inflation mémorielle est surestimée et sa chronologie est différente de celle jusqu'ici considérée et qui se base d'abord sur la production historiographique et les commémorations officielles. De plus, la chronologie de l'investissement de l'Etat et des pouvoirs publics doit être mise en rapport avec celle de l'évolution associative pour nuancer le rapport de causalité le plus souvent mis en avant. L'Etat n'est pas suiveur mais bien prescripteur en matière de constitution de la mémoire en question sociale.

2. **Appréhension erronée des acteurs-clés du phénomène mémoriel**
Sur un plan qualitatif ensuite, S. Gensburger émet l'hypothèse que les acteurs sociaux qui se sont engagés depuis les années 70 sur la "mémoire" ne sont pas seulement que ceux, ni même peut-être principalement, qui se rattachent aux fameuses identités particulières : "antillais", "juifs", "descendants d'immigrés" ou "bretons" etc. Ce n'est pas forcément sur la "mémoire de la Shoah" ou sur celle de "l'esclavage", par exemple, que la thématique mémorielle a émergé en premier.
A cet égard, il est important de saisir comment elle a pu faire son apparition dans les associations d'anciens combattants puisque c'est dans ce ministère qu'elle a émergé en premier.
Par ailleurs, ayant constaté une tension entre administration centrale des anciens combattants et celle de la culture sur la question de la constitution de la "mémoire" en catégorie d'action pour l'Etat, S. Gensburger avance que cette tension et cette dichotomie se retrouvent à l'échelle des associations.

3. **Trajectoire des acteurs associatifs : la mémoire comme lieu de réinvestissement de luttes antérieures**

Bien que cet aspect du projet de recherche ne relève pas du périmètre de notre travail, nous l'exposons néanmoins afin d'en donner un aperçu global. Il est proposé de procéder à une analyse prosopographique des acteurs associatifs qui s'investissent dans le champ de la mémoire. Cette analyse entend montrer que les acteurs militant dans le domaine de la mémoire proviendraient en fait d'autres secteurs de mobilisation. La mémoire, une fois constituée en champ légitime (et socialement intéressant) par l'Etat, serait ainsi devenue un lieu de réinvestissement de luttes passées. Afin de vérifier cette hypothèse, une étude des trajectoires des acteurs associatifs est donc prévue.

1.3 Phénomène mémoriel et analyse de données textuelles

Comme nous venons de le voir, le projet dans lequel nous nous situons entend étudier finement l'apparition et le développement du phénomène mémoriel, en France. Cela passe par l'analyse des acteurs ainsi que celle de leur vocabulaire, dans le temps et dans l'espace. L'hypothèse d'un rôle clé joué par le ministère des Anciens combattants (et en lien avec lui, des associations d'anciens combattants) dans le phénomène mémoriel explique le choix, comme source des données à traiter au moyen d'outils lexicométriques, des déclarations d'associations au *Journal officiel*. L'étude du lexique de ces déclarations devrait en effet permettre de mesurer l'importance (ou non) du phénomène mémoriel au niveau de la demande sociale mais aussi d'en voir les éventuelles évolutions temporelles et les répartitions géographiques. Les déclarations d'associations constituent ainsi un moyen d'approcher un phénomène social (la mémoire), via son vocabulaire, lui-même phénomène social.

C'est pourquoi le recours à des outils d'analyse de données textuelles a été décidé dans le cadre de ce projet : outre la masse de données qu'ils permettent de traiter, ces outils peuvent également offrir des pistes interprétatives intéressantes pour la recherche. Car les matériaux textuels analysés constituent autant d'indices de phénomènes sociaux, qui peuvent alors être appréhendés au moyen d'une analyse lexicométrique.

Afin de ne retenir que les déclarations pertinentes pour l'analyse parmi l'ensemble des déclarations publiées au *Journal officiel*, un filtrage par mots-clés mémoriels ("mémoire", "commémoration", "anciens combattants", "souvenir", "célébration") a été appliqué aux données du JO. Ce choix s'explique par le fait que les termes de "commémoration", puis de "mémoire" (et évidemment celui de "anciens combattants"), sont inscrits dans les missions du ministère des Anciens combattants, comme vu plus haut. Ont été ajoutés à ces premiers termes celui de "souvenir", qui provient de la tradition des anciens combattants et celui de "célébration", qui s'enracine dans celle du ministère de la Culture. Enfin, pour compléter l'appréhension du « phénomène mémoriel » et multiplier les angles par lesquels l'aborder, un dernier terme s'est vu adjoint aux cinq mots-clés initiaux : celui de "patrimoine", qui permet de capturer une dimension mémorielle absente des autres registres (biens matériels et immatériels à conserver, protéger et transmettre).

La période de référence retenue pour l'analyse débute après-guerre (1945) jusqu'à nos jours. Il s'agit en effet d'analyser sur le long terme l'évolution lexicale des déclarations, sans préjuger de périodes plus propices que d'autres à vérifier les hypothèses de recherche.

L'objectif de cette étude lexicométrique du vocabulaire des associations est de permettre de délimiter des ordres de grandeurs (évaluation quantitative) et d'analyser les distributions du phénomène (répartition chronologiques, territoriale, thématique, etc.).

Nous allons désormais détailler les questions qui ont guidé l'analyse lexicométrique des déclarations d'associations au JO.

1.3.1 Analyse lexicale

Une attention sera portée aux questions de glissements entre lexiques. Il s'agira d'analyser les éventuels recoupements thématiques qui peuvent exister au sein des déclarations d'associations. Nous donnons ci-après un aperçu des questions qui ont guidé notre étude :
Quels sont les termes mémoriels les plus employés ? Ceux qui émergent dans le temps ? Des groupements lexicaux particuliers sont-ils constatés ? Et comment évoluent-ils ? Des thèmes régressent-ils dans le temps ? Au profit de quels autres ? Quels sont les termes qui se substituent aux anciens ? Des relations spécifiques se nouent-elles entre certains mots-clés ? Vérifie-t-on les liens évoqués plus haut, entre d'un côté "mémoire" et "commémoration" et de l'autre, "anciens combattants" et "souvenir" ? Sinon, sont-ils davantage reliés à d'autres termes ? Ces regroupements lexicaux privilégiés se transforment-ils dans le temps ? Dans l'espace ?
Quelles sont les premières associations à mobiliser le terme "mémoire" ? En lien à quels thèmes ? Des thèmes sont-ils dominants et cette centralité varie-t-elle selon les périodes et les départements ? Des thèmes sont-ils spécifiques à certains départements ? D'autres thèmes seraient-ils uniformément répartis sur l'ensemble du territoire ?
Que peut-on dire des déclarations contenant des termes « communautaristes » ? Emploient-elles des termes mémoriels ? Si oui, lesquels ? Peut-on voir se dégager une "mémoire juive", une "mémoire de l'esclavage", une " mémoire bretonne", etc. ?
Les expressions de « devoir de mémoire », « lieu de mémoire », « questions mémorielles » ou « lois mémorielles se retrouvent-elles dans le vocabulaire des associations ? Si oui, peut-on en dater l'avènement ? Quels types d'associations sont-ils concernés ?
En parallèle de l'analyse thématique, une analyse de la répartition temporelle et spatiale des associations devra donc être menée, que nous détaillons ci-dessous.

1.3.2 Analyse chronologique

- Pour chaque mot-clé : comment les déclarations contenant tel mot-clé se distribuent dans le temps ? Une année est-elle davantage caractérisée par l'enregistrement de déclarations contenant tel mot-clé ?
- En se fondant sur les liens présentés plus haut entre certains vocables et les ministères des Anciens combattants et de la Culture : analyser la distribution temporelle des déclarations contenant conjointement les mots-clés "mémoire" et "commémoration", de même avec "anciens combattants" et "souvenir".
- Pour l'ensemble des mots-clés, hors celui de « anciens combattants ».
- Pour le mot-clé « anciens combattants » seul.
- Pour l'ensemble des mots-clés, pris conjointement.

Cette analyse devrait permettre de faire apparaître, par exemple, de manière différentielle, la date à laquelle le nombre de création d'association contenant le mot-clé "mémoire" est supérieur à celui des déclarations contenant "célébration", etc.
La prise en compte de certaines dates charnières, d'un point de vue institutionnel, politique, social (année de création de journées nationales de commémoration, controverses publiques, etc.), peut également être introduite dans l'analyse, afin de voir si ces événements influencent (ou non) l'orientation lexicale des associations ou si, au contraire, la création d'associations contenant certains termes en lien à ces événements les précèdent.

1.3.3 Analyse de la répartition géographique

Le même type de questions se pose sur la localisation géographique des associations. Il s'agira d'établir une analyse de la distribution spatiale des associations, en fonction de la mention (ou non) de tel ou tel mot-clé dans leur déclaration ou de la conjonction de deux mots-clés (ou plus). De cette analyse, peut-on voir se dégager des départements plus « mémoriels » que d'autres ? Ou des départements qui seraient davantage le lieu d'exercice d'associations d'anciens combattants (en lien à des lieux de conflits passés, par exemple) ?

Afin d'affiner l'analyse, une approche croisée entre distribution temporelle et géographique peut être envisagée : les départements présentant davantage de déclarations « mémorielles » que d'autres concernent-ils une certaine période ?

1.3.4 « Effets d'intéressement » des associations ?

L'une des finalités des analyses des déclarations d'associations est d'étudier les possibles effets d'intéressement que susciterait le phénomène mémoriel : Est-ce que le phénomène d'institutionnalisation d'un objet social (la « mémoire ») crée des effets d'intéressement auprès des associations ? Autrement dit, peut-on percevoir, au moyen de l'analyse lexicale, une appropriation du vocabulaire mémoriel suite à la « légitimation » de l'objet par l'action publique ?

Cette question des « effets d'intéressement » peut être traitée de plusieurs façons :
- A partir de l'analyse des modifications (titre et objet) de déclarations : Peut-on constater l'ajout du terme « mémoire » (ou l'un des autres mots-clés mémoriels) dans les modifications de déclarations ? Ces modifications ont-elles davantage lieu à certaines dates ? Dans certains départements ? D'autres modifications lexicales sont-elles concomitantes ? Lesquelles ? A quels termes les mots « mémoriels » se substituent-ils ?
- A partir des fichiers de données fournis par les municipalités (hors JO) : ces fichiers constituent le reflet des associations en contact avec les villes. Les fichiers de la ville de Paris, par exemple, fait état de domaines d'activité des associations (dont le champ « mémoire » notamment). Une analyse du lexique de ces sources peut être réalisée, à comparer à celle effectuée avec les données du JO (sur les mêmes zones géographiques). De cette manière, il sera possible de répondre aux questions suivantes : Des thématiques sont-elles privilégiées par les associations dans certaines municipalités ? Si oui, à quelle date ont été créées ces associations ? Ces informations correspondent-elles à celles issues des données du JO ?
Cette analyse devra être reliée à la manière dont les municipalités (via l'étude notamment des organigrammes et des fonctions dédiées à la mémoire) labellisent la mémoire (« mémoire et monde combattant » ou « mémoire et culture », ou encore, « mémoire et droits de l'homme », etc.).

Si de tels effets d'intéressement sont révélés par l'analyse lexicale des déclarations, le projet de recherche procédera alors à une analyse du contexte, institutionnel et social notamment, pour voir si cette orientation de l'activité associative relève d'un intérêt, notamment financier (au travers de subventions), des associations.

Les parties 2 et 3 de ce mémoire seront consacrées à l'exposé de l'analyse des données textuelles que nous avons opérée de la question mémorielle : la méthode employée et les résultats obtenus avec les deux logiciels lexicométriques, Calliope et Alceste seront décrits mais également les limites de ce travail ainsi que les pistes de réflexion proposées dans le cadre d'un prolongement du projet de recherche.

2 Les humanités numériques

L'objet de cette partie est de présenter le domaine dans lequel se situe notre projet d'analyse de données de la question mémorielle : les humanités numériques, et plus particulièrement la lexicométrie Après l'exposé des spécificités et des enjeux sous-jacents qui caractérisent les humanités numériques, nous nous attacherons à décrire l'analyse de données textuelles, en en montrant les intérêts mais également les limites – qui impliquent d'adopter un ensemble de précautions pour être déjouées. Nous terminerons par un panorama des divers outils d'analyse de données textuelles, dont une présentation du mode de fonctionnement des deux outils sélectionnés pour notre travail, Calliope et Alceste.

2.1 Qu'est-ce que les humanités numériques ?

Qu'est-ce les humanités numériques ? Question complexe à laquelle nous n'aurons pas la prétention de répondre, tant les débats sont nombreux et les positions divergentes. Un socle commun semble néanmoins faire consensus : les humanités numériques sont ce « domaine » situé à la croisée des technologies numériques et des sciences humaines et sociales. Mais la délimitation et la définition de cette intersection ne s'imposent pas d'emblée… Sont-elles un ensemble de disciplines œuvrant collectivement ou une transdiscipline ? Permettent-elles l'élaboration de nouveaux objets de recherche, de nouvelles problématiques et de nouvelles méthodes ? Assistons-nous à un bouleversement épistémologique et méthodologique ? Ou représentent-elles, plus simplement, l'adaptation de méthodes et de pratiques plus anciennes aux nouveaux usages du numérique ?
Il ne s'agit pas ici de proposer une définition des humanités numériques mais plus prosaïquement de pointer quelques-unes des questions que ce nouveau champ soulève ainsi que les enjeux qui y sont impliqués. Quoi qu'il en soit, notons simplement que toutes les disciplines des sciences humaines et sociales sont plus ou moins concernées par cette transformation profonde (la littérature, la philologie, l'histoire et la linguistique en premier chef – en raison de leur accès privilégié à de grands corpus de documents –, mais aussi, l'archéologie, l'art et l »histoire de l'art, la géographie, la sociologie, etc.), ainsi que tous les supports d'information et types de données (texte, image, son, document audiovisuel, données géographiques, numériques, etc.), que le recours à des technologies « computationnelles » et de visualisation (lexicométrie, fouille de textes, data mining, analyse et visualisation de réseaux, système d'information géographique, etc.) se répand de plus en plus dans les milieux académiques.
Sans savoir ce qu'elles sont, nous pouvons déjà dire qu'elles sont très présentes. Signe d'une intense activité scientifique, l'on ne compte plus les projets de recherche, colloques, journées d'étude, publications, etc. qui s'inscrivent dans ce « domaine ». « A new critical mass of digital work represented at major conferences like the Modern Language Association and the American Historical Association; new recognition of the need for standards for evaluating digital work for tenure and promotion; new digital humanities centers cropping up like mushrooms, with concomitant digital humanities cluster hires; the words "and digital humanities" suddenly ubiquitously tacked onto job ads; new grant opportunities; a proliferation of THATCamps. [9, CECIRE] »
Sujet de débat dans les milieux académiques, l'usage du numérique et la modification concomitante de la matière sociale constituent aussi à un sujet d'actualité, une question sociale. Preuve en est le récent projet de loi pour une République numérique[5], qui a suscité

[5] En cours d'élaboration, ce projet de loi entend définir la République numérique. « Le numérique et ses usages sont au cœur d'un vaste mouvement de transformation de notre économie, de redéfinition de nos espaces publics et privés, et de construction du lien social. Les conséquences de ces évolutions sont dès à présent globales, et dessinent l'avenir de l'ensemble de notre société. La République du 21e siècle sera nécessairement numérique : elle doit anticiper les changements à l'œuvre, en saisir

de nombreuses contributions de praticiens des humanités numériques. Le numérique et ses pratiques seraient-ils les gisements de richesse du XXI[e] siècle ? Un nouvel eldorado ? L'ensemble de la société semble de fait concernée par ces transformations majeures qui affectent aussi bien les manières de créer que les liens sociaux.
Conscientes des enjeux (sociaux, économiques, scientifiques, politiques, institutionnels, etc.) qu'impliquent l'avènement et la propagation du numérique dans nos modes de vie et de penser, les digital humanities (DH) s'efforcent, depuis les années 70, d'en promouvoir la diffusion, par la création d'associations nationales et internationales[6].

2.1.1 Une difficile définition

Actives, les humanités numériques n'en demandent pas moins à être circonscrites. Comment les présenter ? En 2004, S. Hockey insistait sur le fait que « Tracing the history of any interdisciplinary academic area of activity raises a number of basic questions. What should be the scope of the area? Is there overlap with related areas, which has impacted on the development of the activity? What has been the impact on other, perhaps more traditional, disciplines? Does a straightforward chronological account do justice to the development of the activity? Might there be digressions from this, which could lead us into hitherto unexplored avenues? Each of these questions could form the basis of an essay in itself. [22, SCHREIBMAN et al.] » De même, la page Wikipedia des Digital humanities précise que « The definition of the "digital humanities" is something that is being continually formulated by scholars and practitioners; they ask questions and demonstrate through projects and collaborations with others.[7] »
Des éléments de réponse peuvent néanmoins être trouvés dans le texte du Manifeste des digital humanities[8], rédigé en 2010, qui définit les humanités numériques de la manière suivante :
« 1. Le tournant numérique pris par la société modifie et interroge les conditions de production et de diffusion des savoirs.
2. Pour nous, les digital humanities concernent l'ensemble des Sciences humaines et sociales, des Arts et des Lettres. Les digital humanities ne font pas table rase du passé. Elles s'appuient, au contraire, sur l'ensemble des paradigmes, savoir-faire et connaissances propres à ces disciplines, tout en mobilisant les outils et les perspectives singulières du champ du numérique.
3. Les digital humanities désignent une transdiscipline, porteuse des méthodes, des dispositifs et des perspectives heuristiques liés au numérique dans le domaine des Sciences humaines et sociales. »

Cette définition, large, des humanités numériques peut être interprétée de deux manières : source de richesse, suscitant réflexions et inventions ou signe d'un manque de cohérence. Selon P. Mounier, « Melissa Terras critique ce qu'elle appelle la logique du « chapiteau » (« big tent » en anglais, titre qui avait été choisi pour la conférence « Digital Humanities » organisée la même année à Stanford) : cette logique « inclusive », qui accueille tout le monde pourvu que l'on puisse justifier d'une intersection quelconque entre une des disciplines des sciences humaines et sociales et les technologies numériques, pose aux humanités numériques la question de leur manque de cohérence. Comment établir des critères de qualité, comment établir une offre de formation, délivrer des diplômes, évaluer des publications lorsque l'hétérogénéité est aussi grande à l'intérieur du champ ? [18,

---

pleinement les opportunités, et dessiner une société conforme à ses principes de liberté, d'égalité et de fraternité. » Voir la plateforme dédiée au projet : https://www.republique-numerique.fr.
[6] Voir la liste d'associations nationales et internationales, notamment sur le site d'ADHO (Alliance of Digital Humanities Organizations, http://www.adho.org). A noter la création en 2014 de l'association francophone des humanités numériques, Humanistica : http://www.humanisti.ca. Pour une présentation, sous forme de mindmap, des divers réseaux structurant les DH, voir : https://www.mindmeister.com/fr/84377235/digital-humanities
[7] Page Wikipedia « Digital humanities » : https://en.wikipedia.org/wiki/Digital_humanities
[8] Manifeste des digital humanities : http://tcp.hypotheses.org/318

MOUNIER] » Mais ce caractère flou, non fixé n'est-il pas précisément l'indice d'un mouvement en cours de constitution, de méthodes en cours de renouvellement ?

2.1.2 « Une question de lexique » [9]

Cette absence de définition précise se double – à moins que cela ne constitue justement que l'autre face de la même question – d'une difficulté à désigner les « humanités numériques », à les nommer, notamment en français. Le flottement du signifiant, dans la traduction française de l'expression anglaise, ne renvoie-t-il pas à la délimitation problématique du signifié ? Pourquoi avoir fait le choix du terme « humanités » ? Est-ce la volonté de lier présent et passé académiques au sein d'une lignée historique qui en dirait la tradition ? Est-ce l'inscription initiale des humanités numériques dans l'analyse philologique qui transparaît ? N'est-ce pas réducteur, n'est-ce pas obsolète ? Ce terme pointerait-il davantage vers une humanité dont il s'agirait de rappeler le primat ? Une humanité non seulement à replacer au centre des débats mais aussi de laquelle la science doit se rapprocher, un humanisme du XXI^{e} siècle en quelque sorte ? Le terme « humanités » permet-il de sortir des clivages et problèmes de recouvrement disciplinaires que connaissent les « sciences humaines et sociales », aux niveaux institutionnels et culturels ?
De même, ces recherches sont-elles « numériques » ou « digitales » ? Faut-il mettre l'accent sur la dimension computationnelle, informatique ou sur le « faire avec les doigts », que traduirait « digitales » ? C. Clivaz[10] revendique cette désignation, plus proche selon elle, de l'« essence » des humanités : « Je soulignerai simplement ici qu'"ordinateur" ou l'anglais "computer" désignent un concept cérébral. Or, d'une part, avec la culture de l'Iphone / Ipad, nous entrons en contact avec le monde digital avec nos doigts. […] D'autre part, le chercheur en Humanités Digitales tient de l'Homo Faber et réellement fabrique, crée les nouvelles sciences humaines et sociales. C'est de ces nouveaux moyens d'expression culturelle et académique que naissent en un temps second les questions de recherche fondamentales, complètement transformées. [12, CLIVAZ] »

2.1.3 Réhabilitation de l'Homo Faber

L'essence des humanités numériques résiderait donc dans un faire.
La page Wikipedia (en français) des humanités numériques les présente « comme l'application du "savoir-faire des technologies de l'information [et de l'informatique / infosciences] aux questions de sciences humaines et sociales".[11] » Or, ce rapprochement de savoir-faire technologiques de questions scientifiques ne doit cependant pas être conçu sous le mode d'une simple adjonction, l'apposition de pratiques informatisées à des méthodes qui leur seraient extérieures et indépendantes. En un mot, les humanités numériques ne sont pas des « sciences » plus du « code ». Comme le formule justement N. Cecire, « now that digital humanities is in vogue, there is an overwhelming temptation to believe that the academia problem has at last been solved through the New Criticism plus code. It's the "plus" that makes Hughes's comment so devastating: he puts his finger on a merely paratactic, additive concatenation that is the impoverished version of what can and should be a much more paradigmatic change. In other words, it should not be possible to have the "plus" without the two terms—"digital" and "humanities"—themselves changing. [9, CECIRE] ».
Sans prétendre définir ce que ce changement serait, disons simplement que l'introduction de nouvelles techniques au sein des recherches en sciences humaines et sociales ne se réduit donc pas à la manipulation de fichiers numérisés ou numériques, là où auparavant régnait le document papier. Les humanités numériques ne résument pas à l'avènement d'une

[9] Nous empruntons ici le titre de l'article de [24, URBANIAK] qui présente un panorama de cette question.
[10] C. Clivaz dirige le Digital Enhanced Learning at the Swiss Institute of Bioinformatics (VITAL-IT, Lausanne, CH).
[11] Page Wikipedia « Humanités numériques » : https://fr.wikipedia.org/wiki/Humanités_numériques

transformation de support ni même à l'introduction d'algorithmes au sein de la méthode scientifique.
Car le faire à l'œuvre dans les humanités numériques relève d'un savoir incarné en pratiques, qui redéfinit les « relations que les disciplines des sciences humaines et sociales établissent avec la technologie au sein du processus de recherche. Et de fait, plusieurs observateurs ont remarqué que les humanités numériques ne se présentent pas d'emblée comme une « théorie » ou une nouvelle épistémologie dans le champ scientifique, mais sont concentrées historiquement sur des questions d'outils et de méthodes, dans la solution de problèmes pratiques et la fabrication d'outils pour la recherche. « More hack, less yack » (littéralement : plus de piratage, moins de bla-bla), affiche avec effronterie le Center for History and New Media. […] Ce mot d'ordre révèle une orientation importante du domaine, davantage tourné vers le « faire » (make) que vers le discours, et dont les productions sont plus souvent des outils et des dispositifs que des publications. [18, MOUNIER] »
Ce faire, redécouvert ou plutôt revalorisé par les humanités numériques, n'est que la réhabilitation récente de la réalité du processus qui meut toute production intellectuelle, la promotion d'une réflexivité proprement non discursive. En cela, elles luttent contre une représentation tronquée qui voudrait dissocier le travail de l'intellect de tout lien à la sphère pratique, distinguer (et opposer) œuvres de l'esprit de celles des mains. A. Berra, dans un article au titre évocateur, nous rappelle justement que « les pratiques culturelles sont profondément inscrites dans l'histoire des techniques. Il est urgent de prendre conscience toujours davantage du maniement des objets, contre une naturalisation qui occulte les processus de production et de diffusion. L'écriture elle-même est une technologie, comme l'ont soutenu Walter Ong ou Roger Chartier aussi bien que les théoriciens de l'hypertexte. Ce qui peut paraître le plus mécanique, le plus froid, le plus étranger est tout à fait analogue sur ce plan à d'autres technologies, qui ne paraissent plus naturelles ou plus immédiates que par un effet de l'habitude. [6, BERRA] » L'introduction de nouvelles techniques a donc permis de redécouvrir que le penser ne peut être dissocié d'un faire, et que faire, c'est déjà penser.

2.1.4 Un faire ensemble

Cette revalorisation du faire constitue l'un des pendants d'une nouvelle définition du travail scientifique comme œuvre collective, fondée sur une distribution des savoirs et savoir-faire. La collégialité et la collaboration qui caractérisent les projets au sein des humanités numériques sont tout autant imposées que revendiquées. Imposées dans la mesure où nul ne peut prétendre maîtriser seul l'ensemble des connaissances et compétences que réclament ce type de projet mais aussi revendiquées en tant que valeurs proprement éthiques devant présider à une nouvelle organisation du travail scientifique mais, plus fondamentalement et plus largement, à une nouvelle organisation des relations sociales. Ce combat est bel et bien un combat politique, selon P. Mounier, qui rappelle que « le manifeste de 2010 rattachait les humanités numériques à un ensemble de valeurs dont il assurait la promotion. Pour cette raison, il se situait dans la continuité de tout un ensemble de textes caractéristiques de ce champ : en plus de se définir et de tracer les frontières de leur domaine, nombre d'acteurs des humanités numériques tentent, dans de multiples textes, de définir les valeurs auxquelles ils rattachent leurs pratiques. Cette particularité est étonnante car elle est en rupture avec la situation de la plupart des disciplines des sciences humaines et sociales. Ces dernières, en effet, définissent aujourd'hui un domaine d'activité professionnel qui s'interroge souvent sur ses méthodes, quelquefois sur la place que lui attribue la division sociale du travail, très rarement sur ses finalités. Absorbées par la cité scientifique, les humanités et les sciences sociales sont supposées en adopter les mœurs et les normes sans que cela fasse l'objet de discussions particulières. Quelles sont donc ces « valeurs » que les humanités numériques revendiquent comme leur appartenant en propre et qui les singularisent ? "Openness, collaboration, collegiality, connectedness, diversity, experimentation" [18, MOUNIER] ».
Cette volonté de (ré)inscription de la science dans la vie de la cité se traduit également par la volonté d'œuvrer au bien commun que soulignait déjà le point 7 du Manifeste des digital

humanities : « Nous avons pour objectifs le progrès de la connaissance […] et l'enrichissement du savoir et du patrimoine collectif, au-delà de la seule sphère académique[12] ». Perspective réaliste ou utopique ?

2.1.5 Les humanités numériques, le cheval de Troie d'un nouvel esprit du capitalisme ?

Déjà peut-on remarquer que l'engouement dont jouissent les humanités numériques auprès des instances tant étatiques, institutionnelles qu'économiques en nuance peut-être la capacité subversive. Leurs modes d'organisation semblent, de plus, tout à fait adaptées aux caractéristiques contemporaines de l'organisation économique. « L'organisation par projet par exemple, qui permet de réunir des équipes hybrides et temporaires autour de la réalisation d'un dispositif particulier. Le développement des humanités numériques est donc au minimum l'une des manifestations de ce mouvement plus vaste de redéfinition et de concurrence des modèles d'organisation au sein des institutions scientifiques et culturelles. Les humanités numériques participent-elles de ce mouvement général ? Constituent-elles le cheval de Troie d'une société en réseau, du nouvel esprit du capitalisme et du new public management au sein d'un secteur structuré selon une tradition académique séculaire ? " [18, MOUNIER] » Cette adaptation des humanités numériques aux modes d'organisation économiques transparaît dans les termes mêmes en usage dans le domaine. « "Hands-on," "getting your hands dirty," "digging," "mining," "building"—these terms offer quite a specific vision of what constitutes doing, conjuring up economic productivity of a distinctly social, distinctly virtuous, distinctly white, male, blue-collar variety. The (apparently accidental) choice of "building" for the dominant metaphor of digital doing was never an inevitability. The emphasis on collaboration, collegiality, and the like in digital humanities bears striking resemblance to the values of corporate culture (e.g. "teamwork"). The distinctive methodologies of digital humanities are typically represented in comfortingly industrial terms. [10, CECIRE] »
De fait, la revendication d'une distribution plus large et différenciée des tâches semble reproduire la distinction que l'ancienne hiérarchie « savoir vs. technique » pratiquait. Ainsi, remarque N. Cecire, « The epistemology of doing, in a highly collaborative discipline often involving significant division of labor, means that, as labor is distributed across collaborators, so too is the attribution of knowledge. By this I do not mean "credit" so much as epistemological authority. The manual/mental hierarchy, flipped in the valorization of "hack" over "yack", too often returns in full right-side-up force just when it matters for attributing knowledge to the undergraduates hired to scan archival materials, say, or the workers in India who did the base TEI encoding. To espouse collaboration over authorship, one must have an authorial voice to cede; to be "nice," one must be in a position in which "niceness" does not connote "servility." [10, CECIRE] »

Sans vouloir dresser un tableau pessimiste, il est néanmoins important de ne pas occulter les risques potentiels traversant les humanités numériques pour les discuter, les critiquer et interroger le rôle qu'elles peuvent jouer au sein de leur environnement tant économique, institutionnel, scientifique que politique. Seule l'évaluation constante des promesses visées par les humanités numériques à l'aune des risques potentiellement encourus peut aider à circonscrire ces derniers pour tenter d'atteindre les premières.

2.1.6 Les humanités numériques, un mouvement aux origines multiples

Pour mieux saisir ce que sont les humanités numériques, nous proposons un détour par l'histoire de ce mouvement. Mais quels jalons retenir pour présenter une telle histoire ? Où en situer le commencement ? Si l'aspect instrumental lui est essentiel, son histoire se

[12] Manifeste des digital humanities : http://tcp.hypotheses.org/318

confond-elle avec l'histoire des techniques ? Faut-il remonter à l'invention de l'écriture, voire du premier outil ? Autant de questions auxquelles il est difficile de répondre. Ainsi, A. Berra remarque-t-il que, « qui veut faire l'histoire des humanités numériques est d'emblée confronté à un véritable problème. Nous pouvons lire les réflexions d'acteurs pionniers, qui se demandent quels seraient les critères pour délimiter cette histoire. Faut-il commencer avec l'imagination antique et les automates de l'Iliade, quand Homère imagine que les instruments d'Héphaïstos s'agitent seuls comme autant de proto-robots ? Faut-il commencer avec les techniques de calcul, qui ont une histoire autonome dans divers champs de savoir ? Faut-il commencer avec l'ordinateur, comme nous aurions évidemment tendance à le faire ? Comment mêler ces différentes histoires et leur contribution aux « humanités » ? Quels sont les documents dont nous disposons ? [6, BERRA] ».

Par commodité, nous prendrons comme point de départ des humanités numériques le développement et l'utilisation, après-guerre, des premières calculatrices électroniques (ordinateurs), appliquées notamment aux matériaux textuels. Le recours à des méthodes informatisées, computationnelles s'enracinent de fait dans de nombreux courants de recherche et types d'analyse (analyse philologique, analyse linguistique, statistique textuelle, statistique lexicale, analyse de contenu, analyse de discours, intelligence artificielle, linguistique computationnelle, etc.), aux finalités variées (recherche de concordances, corrections orthographiques et syntaxiques, description de textes, élaboration de résumés, de traductions, de typologies de documents, exploration de données textuelles, extraction de connaissances, etc.).
Dans le domaine des études littéraires et philologiques, il est conventionnellement admis de faire débuter les humanités numériques en 1949, avec le travail pionnier de Roberto Busa sur les techniques de concordance automatique appliquées à l'œuvre de Saint Thomas d'Aquin. Ainsi, « Roberto Busa souhaitait étudier le vocabulaire de la présence, de l'incarnation, dans les œuvres de Thomas d'Aquin, ce qu'il ne pouvait pas faire en se contentant de rechercher des occurrences de substantifs. Il devait aussi, dans les millions de termes de son corpus, s'intéresser à un mot comme la particule latine *in*, « dans ». La compilation et l'usage d'un index complet constituaient un travail surhumain. La lecture intensive traditionnelle ne pouvait pas suffire. La concordance automatique était la solution. Il a cependant fallu un travail de trente ans pour mener à bien ce projet [6, BERRA] ».
Représentatives de l'essor des techniques informatisées aux sein des sciences humaines, les années 60 voient alors paraître le premier numéro de *Computers and the Humanities*. Le vocable humanités numériques prend aujourd'hui le relais de ce que l'on nommait à l'époque « Humanities Computing », à savoir les sciences humaines assistées par l'ordinateur (recours aux outils informatiques dans l'analyse de corpus, notamment pour l'étude des textes littéraires anglais). Ce traitement informatisé de corpus littéraires constitue l'une des origines de la statistique lexicale avec, en France, la saisie informatique dès le début des années 60 d'œuvres littéraires des XIX^{e} et XX^{e} siècles au sein de la base Frantext[13], dans le but de servir de support de documentation et de base d'exemples littéraires au dictionnaire *Trésor de la langue française*. Pour la rédaction de ce dictionnaire, il a en effet fallu rassembler, analyser et traiter avec l'aide de l'informatique d'immenses fonds documentaires bibliographiques, lexicologiques et surtout textuels. Cette initiative s'est poursuivie pour accueillir d'autres types de textes (scientifiques et techniques), couvrir une plus grande période temporelle (du XVI^{e} au XX^{e} siècles) et développer les techniques linguistiques associées (catégorisation grammaticale, lemmatisation, requêtes via expressions régulières, par séquences de termes, etc.). L'association de l'analyse de corpus textuels et de méthodes

[13] Voir la base Frantext : http://www.frantext.fr. Initialement développée par l'Institut national de la langue française (CNRS), cette base porte la marque des grands programmes lexicographiques développés par le CNRS depuis 50 ans. Elle est aujourd'hui gérée conjointement par l'ATILF (Analyse et traitement de la langue française, CNRS, Nancy Université) et le CNTRL (Centre national de ressources textuelles et lexicales, CNRS), qui met à disposition sur son portail un ensemble d'autres ressources textuelles et lexicales informatisées et d'outils permettant un accès intelligent à leur contenu : http://www.cnrtl.fr.

statistiques (notamment multidimensionnelles) donne naissance à l'analyse des données textuelles, que nous détaillerons plus loin.
D'autres origines aux humanités numériques que celle des sciences humaines peuvent également être citées, concernant notamment les techniques de traduction automatique, dont l'idée apparaît dans les années 40. Ces recherches s'inscrivent dans un contexte de guerre froide et de concurrence. Mais, outre des intérêts strictement militaires et politiques liés à cette période, la traduction d'articles russes, particulièrement dans le domaine spatial, était jugée indispensable par les scientifiques américains. Ainsi « en janvier 1954 eut lieu à New York la première démonstration sur ordinateur, une machine IBM 701, qui déclencha une accélération des recherches. Il s'agissait de la traduction de russe en anglais de phrases utilisant un vocabulaire de 250 mots et six règles de syntaxe mises au point par la Georgetown University. Bien que très limitée, cette démonstration fut montée en épingle par la presse et fit grande impression sur le public et certains scientifiques. [15, LEON] » Les recherches en traduction automatique prennent alors, en l'espace d'une dizaine d'années, une dimension mondiale.
A la même époque, les travaux de Wiener (cybernétique) constituent une tentative théorique pour unifier les domaines naissants de l'automatique, de l'électronique et de la théorie mathématique de l'information. De ce courant sont issues les sciences cognitives, l'intelligence artificielle, les thérapies systémiques de l'école de Palo Alto, ou encore les théories biologiques de l'auto-organisation[14].
Certains courants de recherche actuels, dont le traitement automatique des langues, s'inscrivent dans cette double filiation, de traduction automatique et de sciences cognitives.

2.1.7 Les usagers actuels des humanités numériques

Avant d'aborder la lexicométrie, domaine dans lequel se situe notre travail, nous aimerions conclure cette présentation des humanités numériques par un portrait de leurs usagers. Quels sont les praticiens des humanités numériques aujourd'hui ? Nous nous basons sur une enquête « Who are you, Digital Humanists ? », lancée à l'issue du ThatCamp Luxembourg en 2012. « On y découvre une très grande diversité linguistique et géographique, la marginalité de l'anglais comme première langue, mais sa domination comme second idiome. S'y révèlent des Digital Humanities fortement marquées par l'Histoire et les études classiques, mais très peu, beaucoup trop peu, connectées aux disciplines s'intéressant au monde contemporain, d'une part, et au sciences du Web, à la fouille de données, à la fouille de textes, d'autre part. On y découvre également [que] la diversité culturelle est gouvernée par l'Europe et l'Amérique du Nord, et plus précisément par le Royaume-Uni et ses anciennes colonies (Irlande, Canada, Etats-Unis d'Amérique, Australie).[15] »
Nous retiendrons de cette enquête une relative non appropriation de ces méthodes par les sciences sociales (sociologie, sciences politiques et économie, notamment). En effet, semble se profiler une domination des disciplines historiques (paléographie, histoire médiévale, histoire moderne) et des études littéraires, de la philologie, de l'histoire de l'art et de la musicologie A l'opposé, située en fin de classement, apparaît une petite minorité de réponses issues de l'ingénierie des langues et des technologies du web : fouille de textes, sciences du web, traitement automatique de la langue. Cette situation peut s'expliquer par l'origine des humanités numériques (étude des arts et des lettres), ou par le fait que l'enquête n'ait touché que les chercheurs relevant de ces domaines, et que les autres praticiens n'aient pas été enquêtés ou se soient sentis peu concernés. Comme le souligne M. Dacos, il faut noter « l'existence d'un hors-monde qui n'a pas vu l'enquête ou n'y a pas prêté attention. […] De façon très rapide, on pourrait dire que les Digital Humanities d'aujourd'hui sont finalement bien plus humaines que digitales, et bien plus historiques que sociales. »

[14] Voir la page Wikipédia de « Cybernétique » : https://fr.wikipedia.org/wiki/Cybernétique
[15] Pour une présentation détaillée de cette enquête, voir sur Blogo-Numéricus, l'article « La stratégie du sauna finlandais », de Marin Dacos : http://bn.hypotheses.org/11138

3 L'analyse de données textuelles

3.1 Qu'est-ce que l'analyse de données textuelles ?

Tout comme pour les humanités numériques, la terminologie de la lexicométrie présente quelques flottements et plusieurs dénominations coexistent[16] : lexicométrie, statistique lexicale, statistique textuelle, approche quantitative des textes, analyse statistique des données textuelles, qui se voient prolongées par la textométrie et la logométrie (prise en compte des textes et des discours et non du seul lexique). Ces méthodes se situent au croisement de plusieurs disciplines : linguistique, statistique et informatique.
D'autres méthodes de traitement des données textuelles coexistent à côté de la lexicométrie, certaines issues de la linguistique, des sciences cognitives, de l'intelligence artificielle, des neurosciences et de l'informatique notamment (linguistique computationnelle, fouille de texte et text mining, exploration de données, extraction de connaissances, traitement automatiques des langues, linguistique calculatoire, ingénierie des langues), d'autres des méthodes qualitatives. Ces deux types de méthodes, excédant le périmètre de notre étude, ne seront pas détaillés ici.
Disons simplement que les premières méthodes opèrent sur un périmètre plus large que les données textuelles (traitement du signal notamment, avec reconnaissance automatique de la parole, traitement et synthèse de la parole) et avec des ambitions plus importantes : extraire automatiquement des connaissances à partir du traitement des données, et non des représentations schématisées comme le propose la lexicométrie Ainsi, le traitement automatique des langues (TAL), situé à la frontière de la linguistique, de l'intelligence artificielle (représentation de l'information et des connaissances dans des formats interprétables par des machines), de l'informatique et des statistiques, vise à « modéliser et à reproduire, à l'aide de machines, la capacité humaine à produire et à comprendre des énoncés linguistiques dans des buts de communication [25, YVON] ». Les principaux domaines du TAL sont le traitement de la parole, la traduction automatique, la compréhension automatique des textes, la génération automatique de textes, le résumé automatique de textes, la reconnaissance de l'écriture manuscrite, la recherche d'informations et la fouille de texte, la reconnaissance d'entités nommées, l'annotation sémantique, la classification et la catégorisation de documents, etc. D'après l'ATALA (Association pour le traitement automatique des langues), « si la traduction automatique reste une application majeure du TAL, bien d'autres applications ont vu le jour depuis 50 ans, pour former les "industries des langues". Elles s'appuient sur des recherches jetant une passerelle entre la linguistique et l'informatique sans oublier la statistique, fort utile pour extraire des données du gigantesque réservoir de textes disponibles sur le web[17] ».
Le second groupe de méthodes (analyse de contenu), issues de méthodes qualifiées de qualitatives, ne se fondent pas sur des principes statistiques ou, si elles les emploient, ce n'est qu'en guise de contrôle, en fin du processus. Cette démarche consiste à analyser le contenu de documents textuels pour en élaborer une classification, à l'aide de catégories ou de codes attribués par l'analyste aux documents. Ces catégories ou codes peuvent être liées au contenu du document (champs sémantiques) ou au contexte de sa production (source, date, sexe). Des outils désignés par l'acronyme CAQDAS (*Computer assisted / aided qualitative data analysis software*) sont souvent employés dans ce type de méthodes pour faciliter l'annotation par l'expert et la recherche de documents. Nous présenterons ce type d'outils ultérieurement, lors de la présentation des différents outils d'analyse textuelle.

---

[16] Dans ce mémoire, nous utilisons indifféremment les termes d'analyse de données textuelles, de lexicométrie, de statistique textuelle, d'analyse statistique de données textuelles. Pour un glossaire détaillé des termes en usage au sein de la lexicométrie, voir celui de Lebart et Salem : lexicometrica.univ-paris3.fr/livre/st94/STUCGLOS.pdf (*in* LEBART L. et A. SALEM, *Statistique textuelle*, Paris, Dunod, 1994).
[17] Voir le site web de l'ATALA : http://www.atala.org.

3.1.1 Présentation des méthodes implémentées

Nous allons désormais aborder le domaine de la lexicométrie, qui nous concerne ici. Ce domaine est particulièrement vivant en France et depuis 1992, la communauté d'analyse de données textuelles se réunit en congrès tous les deux ans lors de Journées internationales d'analyse statistique de données textuelles (JADT). Les actes sont publiés et accessibles en ligne sur le site Lexicométrica[18]. Comme évoqué plus haut, les premiers corpus étudiés à l'aide de ces méthodes appartenaient plutôt au champ littéraire mais se sont diversifiés dans le temps : étude de textes de partis politiques, de syndicats (notamment au laboratoire de lexicométrie de Saint-Cloud), puis analyses sociologiques, pour désormais couvrir toute une palette de problématiques et de types de documents (incluant l'analyse de tweets ou de commentaires, issus de sources web).

Mais qu'est-ce que la lexicométrie ? Selon M.-A. Polo de Beaulieu, il s'agit d'un ensemble de méthodes de description des textes fondées sur des techniques statistiques. Concrètement, elle permet de « traiter de vastes ensembles de textes (corpus), d'établir leur vocabulaire, de classer les vocables en fonction de leur fréquence, de leur répartition, de leurs catégories grammaticales. Elle établit les contextes d'emploi d'un vocable et les combinaisons les plus fréquentes dans lesquelles il entre. [42, LABBE et LABBE] » Contrairement aux techniques de TAL, les principes statistiques sur lesquels reposent ces outils ne sont « ni linguistiques (ils ne prétendent en aucune façon simuler le fonctionnement de la langue) ni explicatifs (ils ne disent pas le pourquoi des faits) [55, POLO DE BEAULIEU] ».
Les outils de lexicométrie considèrent le texte comme une série de blocs graphiques, pouvant être constitués d'unités (une chaîne de caractères) ou de séquences d'unités (plusieurs chaînes de caractères, délimitées par des espaces ou par des caractères de ponctuation, comme les points et les virgules). Grâce à leurs dictionnaires intégrés, plus ou moins sophistiqués, les logiciels peuvent affecter une catégorie grammaticale à chaque forme graphique, sur laquelle ils se fondent pour opérer, si souhaité, un processus de lemmatisation. Les méthodes statistiques (analyse factorielle des correspondances et classification notamment) sont appliquées sur cet ensemble de formes (réduites ou non), répertoriées en tableaux lexicaux, et en donnent une représentation, schématique, mettant en avant des points saillants, des spécificités lexicales auparavant invisibles, noyées dans la masse des données.
Découper, dénombrer, classer et représenter schématiquement sont donc les maîtres mots de la statistique lexicale.

L'analyse de données textuelles (ADT) s'appuient sur les méthodes de la statistique exploratoire multidimensionnelle [82, LEBART et SALEM], qui met en forme de vastes ensembles de données [textuelles, dans le cas de l'ADT] pour en dégager les structures. « Ces méthodes généralisent la statistique descriptive classique et utilisent des outils mathématiques assez intuitifs, mais plus complexes que les moyennes, variances et coefficients de corrélations empiriques de la statistique descriptive. Les méthodes exploratoires multidimensionnelles recouvrent un grand nombre de techniques qui ont pour objectif de décrire et synthétiser l'information contenue dans de vastes tableaux de données [81, LEBART, PIRON et STEINER] ». Pour ce faire, la statistique multidimensionnelle va calculer des distances (ou des proximités) entre les éléments d'un tableau lexical à double entrée (tableau de fréquences croisées), dans lequel les colonnes représentent des variables (par exemple, chaque sujet interviewé ou une variable de contraste comme l'appartenance socioéconomique ou professionnelle d'un individu, etc.) et les lignes, des individus (les mots, dans le cas d'une analyse de données textuelles, puis chaque regroupement de mots répétés au moins deux fois de façon identique dans le discours). Le premier tableau est donc un tableau de fréquence basé sur le mot et le second sur les segments de phrases répétées.

[18] Voir, pour les JADT, le site http://lexicometrica.univ-paris3.fr/jadt/ et pour la revue *Lexicometrica*, le site : http://lexicometrica.univ-paris3.fr, dont les articles relèvent de domaines plus larges que la seule lexicométrie (lexicométrie, statistique textuelle, linguistiques de corpus, extraction d'informations à partir de corpus de texte, acquisition de connaissances, traitement automatique des langues, etc.).

Dans le cas d'Alceste, qui relève de cette méthode, les colonnes sont constituées de mots et les lignes caractérisent les unités de contexte.
L'ensemble des colonnes permet de définir, à l'aide de mesure de distance, les distances entre lignes et à l'inverse, l'ensemble des lignes permet de mesurer les distances entre colonnes. Un mot (ou plusieurs mots) rapprochent, par exemple, deux individus s'il est (ou s'ils sont) commun(s) aux deux individus. Si tel n'est pas le cas, alors la distance entre les deux individus est considérée comme plus grande.
Or, « s'il est toujours possible de calculer des distances entre les lignes et des distances entre les colonnes d'un tableau, il n'est pas possible de les visualiser de façon immédiate (les représentations géométriques associées impliquant en général des espaces à plus de deux ou trois dimensions) : il est nécessaire de procéder à des transformations et des approximations pour en obtenir une représentation plane. » Cette réduction des informations du tableau lexical, nécessaire, ne doit pas pour autant entraîner de perte significative d'information. Des méthodes statistiques sont donc sollicitées pour opérer cette opération à « moindre frais », qui se subdivisent en deux grandes familles :

3.1.1.1 Les méthodes factorielles

Ces méthodes (dont l'analyse factorielle des correspondances, AFC) consistent à rechercher les directions principales selon lesquelles les points s'écartent le plus du point moyen. « Les méthodes factorielles permettent de gérer simultanément des quantités importantes de données et leur système de corrélations et, par une technique réalisant une sorte de compression, d'en dégager la structure interne, notamment sous forme de graphique-plans. L'objectif est de rechercher des sous-espaces de dimensions réduites (entre trois et dix, par exemple) qui ajustent au mieux le nuage de points-individus et celui des points-variables, de façon à ce que les proximités mesurées dans ces sous-espaces reflètent autant que possible les proximités réelles. On obtient ainsi un espace de représentation, l'espace factoriel, défini par les axes principaux d'inertie et l'on représente les points du nuage dans ce système d'axes. Ces axes réalisent les meilleurs ajustements de l'ensemble des points selon le critère classique des moindres carrés, qui consiste à rendre minimale la somme des carrés des écarts entre les points et les axes. [81, LEBART, PIRON et STEINER] »
L'AFC, diffusée et développée en France par Benzécri, permet notamment de « tester l'indépendance des lignes et des colonnes, mais surtout de décrire comment les données s'éloignent de cette hypothèse en représentant par des "proximités" les associations existantes entre les lignes et les colonnes. [27, BEAUDOUIN] »
Cette méthode est particulièrement adaptée aux tableaux de cooccurrences de termes car elle permet de visualiser et de hiérarchiser l'information à partir du traitement de grandes masses de données. Une mesure de distance entre éléments (mesure du khi2[19]) est sollicitée pour vérifier, notamment, si des structures lexicales sont fortement et systématiquement récurrentes à l'intérieur d'un sous-corpus, au point qu'elles peuvent caractériser le type de discours commun à divers individus.
Ces méthodes permettent donc de synthétiser un volume d'informations important à l'intérieur de plans factoriels (représentations graphiques dans un espace à deux dimensions), fort utiles lorsqu'il s'agit de déterminer la part de discours partagé par une population et, dans un même tableau, ce qui est distinct et spécifique à divers sous-groupes.

3.1.1.2 Les méthodes de classification

Ces méthodes consistent à rechercher des groupes ou classes d'individus qui soient les plus homogènes possibles. Elles procèdent à des regroupements en classes des lignes ou des colonnes, par la recherche de proximités entre les éléments du tableau lexical. Ces méthodes se subdivisent à leur tout en deux, avec d'un côté les méthodes de classification

[19] Le khi2 d'appartenance est une mesure statistique qui vérifie l'interdépendance entre les lignes et les colonnes d'un tableau de contingences (ou tri croisé). Le khi2 mesure la liaison entre une forme (colonne) et son contexte (ligne), ie la probabilité qu'une forme soit par hasard associée à son contexte.

hiérarchique (obtenir une hiérarchie de classes à partir d'un ensemble d'éléments décrits par des variables, ou dont connaît les distances deux à deux) et les méthodes partitionnement (ou classification directe) qui produisent de simples découpages de la population, sans hiérarchie Ces méthodes constituent, selon Lebart et Salem, un bon complément aux AFC car « les AFC permettent avant tout de dégager les grands traits structuraux [des tableaux lexicaux]. Cependant, lorsque le nombre des éléments représentés est important, il devient délicat dans la pratique d'apprécier leurs positions réciproques au vu de seuls résultats graphiques. [82, LEBART et SALEM] »
Nous ne développons pas plus, ici les méthodes de classification hiérarchique car les outils retenus dans le cadre de cette étude se fondant sur elles, elles font l'objet d'une présentation détaillée ultérieure.

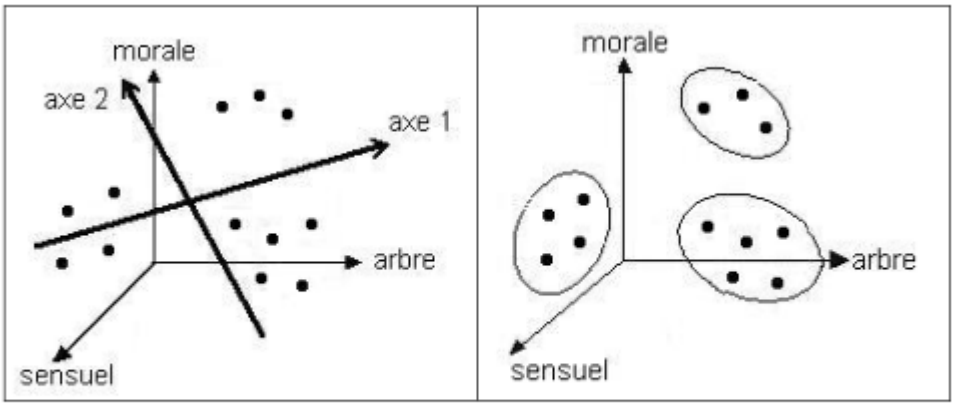


Figure 1 – Deux grandes familles de méthodes de réduction du tableau lexical[20]

Selon V. Beaudouin, on peut distinguer, au sein de la lexicométrie, deux courants différents[21], la statistique lexicale et la statistique textuelle. Le premier courant, issu des travaux de C. Muller à la fin des années 50, compare les données observées dans les textes aux données calculées à partir qu'une référence interne (l'ensemble du corpus) ou d'une référence externe comme le *Trésor de la langue française*. Cette méthode est souvent employée dans l'analyse stylistique des textes littéraires (richesse, spécificité, accroissement et évolution chronologique du vocabulaire notamment). Le logiciel conçu par E. Brunet, Hyperbase, se situe dans ce courant de la statistique lexicale. Il permet notamment de mettre en évidence les termes les plus spécifiques d'une œuvre ou d'un ensemble d'œuvres par rapport à l'ensemble d'un corpus. Le second courant, l'analyse textuelle, s'oppose à la méthode lexicale par son refus de faire appel à une référence externe au corpus. Le principe consiste à se baser sur les seules données du corpus et à les décrire en analysant la distribution de l'information présente dans les données. Cette étude des lois de distribution du vocabulaire dans un corpus provient des travaux de J.-P. Benzécri, dont l'approche elle-même peut être considérée comme dérivée de l'approche distributionnelle du linguiste Z.S. Harris. Il s'agit en effet, pour Benzécri, d'étudier les langues pour en montrer les structures linguistiques. « L'analyse distributionnelle est un ensemble de méthodes formelles pour étudier les langues "de l'extérieur" en éliminant le recours au sens. Le critère général de cette théorie est la position du mot. Les mots ayant la même position dans un contexte identique seront associés à une même catégorie [85, PERETTI DE] » Alceste s'appuie sur cette conception pour établir une classification des énoncés, en fonction de la cooccurrence de leurs mots.

---

[20] Graphique issu de LEBART L., M. PIRON, et J.-F. STEINER, *La sémiométrie□: essai de statistique structurale [En ligne]*, Dunod, 2003.
[21] Pour une présentation détaillée de ces deux courants, voir BEAUDOUIN V., « Statistique textuelle □: une approche empirique du sens à base d'analyse distributionnelle », *Texto [En ligne]*, 2000.

Pour conclure cette présentation succincte de la lexicométrie, il faut souligner que le maître mot de la statistique textuelle est l'homogénéité. En effet, s'il existe trop de différences entre les textes traités (variété des types de documents, des sources, des types d'énoncés, etc.), les méthodes statistiques ne produiront, à la fin du traitement, que des résultats hétérogènes, reflétant les différences initiales. Il faut donc soumettre aux méthodes statistiques des corpus homogènes pour leurs calculs puissent mettre en valeur de micro-différences entre discours, entre documents, ou encore les glissements lexicaux à l'œuvre dans une série temporelle.

3.2 Pourquoi utiliser l'analyse de données textuelles ?

3.2.1 Traiter des corpus volumineux

Comme nous venons de le voir, les méthodes de statistique multidimensionnelle permettent de traiter et de représenter de grandes masses de données, particulièrement adaptées à l'explosion de données qui caractérise notre époque. Ainsi, « pour les années les plus récentes, j'ai une source formidable : le web. Mais comment exploiter cette source qui, à l'échelle d'un historien, est potentiellement infinie ? Comment appréhender une abondance de sources telle qu'il n'est pas possible humainement de la lire intégralement comme j'ai pu lire quelques centaines de boîtes d'archives pour ma thèse ? » se demande F. Clavert [11, CLAVERT] sur son blog, *L'histoire contemporaine à l'ère numérique*.

Là où auparavant les chercheurs devaient consacrer de nombreuses heures à saisir des textes, à les coder puis à attendre les résultats des traitements, ils sont aujourd'hui confrontés à une abondance de documentation, dont les sources se sont multipliées (document papier scanné ou nativement numérique). Pour faire face à cette massification sans précédent de la documentation, le recours à des techniques informatisées semblent s'imposer. Et l'un des intérêts les plus évidents de l'utilisation de logiciels d'analyse de données textuelles est précisément de pouvoir traiter des corpus de données volumineux, impraticable à une lecture / annotation manuelle. Le temps de traitement de tels corpus, largement inférieur à celui que réclame une manipulation manuelle, permet ainsi d'accéder à des corpus auparavant inaccessibles. Par ailleurs, de nouvelles sources (web) ou un accès diversifié à des sources plus anciennes (numérisation d'archives, politique de libre accès aux données publiques) s'offrent à la recherche, notamment en sciences sociales.

Ainsi, « Les logiciels d'analyse de données textuelles « contribuent à desserrer les contraintes de toute sociologie empirique, celles qui sont relatives au volume des matériaux mobilisables dans l'enquête, et ceci à un moment donné où les outils informatiques et les technologies de l'information et de la communication jouent un rôle croissant dans le recueil et l'analyse de données en sciences sociales (Brossaud, Reber, 2006), et où l'internet et les avancées de la numérisation documentaire donnent accès à des matériaux textuels de plus en plus volumineux. […] La constitution de gros corpus textuels diversifiés devient de plus en plus aisée, qu'il s'agisse de documents d'archives, de dossiers de presse, de données « naturelles » enregistrées, de collections d'entretiens[22], de séries documentaires diverses, sans compter l'éventuelle constitution de bases de données d'entretiens ouvertes à l'analyse secondaire[23]. Les logiciels d'analyse textuelle constituent dans ce cadre une ressource pour l'exploitation de ces ensembles volumineux de matériaux. [31, DEMAZIERE et al.] ».

[22] Il n'est que de citer l'initiative BeQuali (Banque d'enquêtes qualitatives en sciences sociales), développée au sein du CDSP (Sciences po / CNRS) dont l'objectif est de donner accès, par une conservation patrimoniale documentée, aux données de la recherche en sciences sociales : http://bequali.fr/fr.

[23] Voir, par exemple, le projet ANR « Réanalyse », qui œuvre à un archivage et à une analyse secondaire des enquêtes qualitatives en sciences sociales : http://reanalyse.hypotheses.org.

3.2.2 Une vision synthétique du corpus

« A la manière des archéologues qui utilisent des vues aériennes de l'espace pour cartographier la région des fouilles leur permettant de découvrir des fragments significatifs d'une vie collective passée, nous sommes en train de circonscrire l'espace de notre corpus lexical et de regrouper des objets et des lieux usuels, avant de tenter d'en donner une description précise et une interprétation fine. [40, KALAMPALIKIS] »
Combinée à la capacité de traiter de gros corpus, l'analyse de données textuelle permet, grâce aux méthodes statistiques implémentées, d'appréhender de manière synthétique l'ensemble des informations contenues au sein d'un corpus – ou plutôt d'en dresser un certain résumé, une certaine vue. Les méthodes de classification ou l'analyse factorielle des correspondances révèlent, rendent visibles des corrélations invisibles à l'œil nu, qui mettent en lumière des points saillants, caractéristiques, parmi une profusion de données. La puissance du chiffre couplée à la force de l'image rendent alors possible une appréhension globale d'une multitude de données qu'aucun traitement manuel n'aurait pu égaler. Comment manuellement décrire les thèmes caractérisant des milliers, voire des millions de documents ?
Cette capacité de traitement et de représentation synthétique se voit renforcée par une mise à distance de la subjectivité de l'analyste. La lecture « manuelle », humaine n'a-t-elle pas tendance en effet, lors de son parcours linéaire des matériaux textuels, à ne mémoriser que ce qui sort de l'ordinaire, que ce qui émerge de l'aplat que constitue le continuum des traces matérielles du texte ? Ne nous rappelons-nous pas davantage l'incident, le pic que le banal, le lisse ? Mais est-ce bien le cas ? Comment apprécier ce pic ? Est-il identique d'un individu à un autre, ou même d'un moment à l'autre ? Les « phénomènes d'habituation à la récurrence, de surcharge cognitive liée au volume de texte [44, LAROSE, KARSENTI et GRENON] » ne doivent-ils pas être pris en compte lors de l'annotation d'un corpus ?
En tenant compte du lisse, précisément, du fréquent, du répétitif, les outils d'analyse de données textuelles sont alors capables non seulement de distinguer ce qui se reproduit mais aussi ce qui se reproduit ensemble, conjointement (cooccurrences) et ce qui ne se produit qu'une fois. En un mot, ces outils séparent le régulier de la rareté mais de manière quantitative, et non plus subjective : des groupes de données, des classes de spécificités lexicales sont alors statistiquement élaborés qui donnent une image quantitativement fondée de leur distribution, de leur structure.

3.2.3 La quantification comme instrument d'objectivation des SHS ?

L'aura dont les méthodes quantitatives sont entourées (du moins, dans certains champs disciplinaires) proviendrait donc de leur capacité essentielle à saisir froidement le « réel », à le résumer sans l'introduction de distorsions humaines, trop humaines. La capacité descriptive des statistiques l'emporterait sur l'interprétation subjective[24]. Ainsi, G. de Peretti évoque la prédominance des sciences de la nature sur les sciences humaines et sociales, des sciences dites « dures » sur des sciences considérées comme « molles ». Et, pour filer la

[24] Nous forçons ici délibérément le trait. Voir notamment, pour une vision plus subtile des modes d'élaboration de la quantification et de ses effets et, DESROSIÈRES A., *La politique des grands nombres. Histoire de la raison statistique*, Paris, Éditions La Découverte, 2010. Ainsi, « La possibilité de manipuler sans les déformer des objets macrosociaux fondés sur des calculs statistiques permet aujourd'hui de circuler sans effort entre plusieurs niveaux de réalités dont les modes de construction sont pourtant fort différents : c'est la magie statistique. De ce point de vue, le rapprochement entre deux types de calculs de moyennes, fortement identiques mais logiquement bien distincts, a été un moment clé : d'une part, la meilleure approximation d'une grandeur de la nature, à partir de mesures différentes issues d'observations imparfaites portant sur un seul objet et, d'autre part, la création d'une réalité nouvelle à partir d'observations portant sur des objets différents mais liés pour l'occasion. Ce rapprochement, rendu célèbre par Quetelet, entre la permanence de l'objet observé plusieurs fois et l'existence de quelque chose de commun entre les objets différents, fournit une solidité toute nouvelle au lien établi alors entre ces objets. »

métaphore, le recours à des méthodes statistiques permettraient à ces sciences molles de se « durcir », de se doter de l'objectivité dont elles seraient intrinsèquement dépourvues. « Le choix [de l'analyse quantitative] repose sur le postulat d'une supériorité des méthodes quantitatives sur les méthodes qualitatives en termes d'objectivité pour traiter cette masse d'information. C'est l'idée de la neutralité des techniques. D'ailleurs, dans le contexte de la méthodologie sociologique, pour certains, ce sont les nombres qui sont la référence implicite du couple "qualitatif/quantitatif" où le terme qualitatif renverrait plutôt à l'absence de quantité […], soit à un mode d'analyse mineur. […] Le postulat sous-jacent est la suprématie des sciences de la nature sur les autres sciences (sociales, politiques ou juridiques) dans la dimension objective. Ceci se retrouve, en particulier, en économie où l'on tente de tout monétariser afin de travailler sur des mesures "réelles". Cette suprématie du quantitatif sur le qualitatif aurait son origine dans les sociétés occidentales des années 1300 où dans, un laps de temps réduit, l'Europe aurait produit sa première horloge mécanique ou la quantification du temps, les premières cartes marines et peintures avec perspective ou la quantification de l'espace et les premières comptabilités à double entrée ou quantification des comptes financiers. [85, PERETTI DE] »
L'introduction de méthodes quantitatives (au moyen notamment de l'analyse lexicométrique) doterait alors les sciences humaines et sociales de critères de scientificité vérifiables, car reproductibles et partant, réfutables.

3.2.4 Un accès privilégié aux matériaux textuels

Pour les chercheurs en sciences sociales, le recours à des outils d'analyse de données textuelles représente un autre atout : celui de donner accès à une masse de documents textuels, phénomènes sociaux s'il en est. Pour les sociologues, historiens, politistes, etc., la conjonction de nouvelles sources, de masses de données et de traitements statistiques constitue l'opportunité d'étudier autrement la matière sociale, de la travailler via d'autres biais, d'autres moyens d'accès. La puissance de calcul, alliée au processus d'objectivation des démarches et à la quantité de données, ouvrent ainsi des perspectives d'exploration inédites. Les matériaux mêmes, rassemblés au sein de corpus textuels, représentent d'inestimables ressources disponibles dont le chercheur peut extraire sens et relations. Car concentrées en leur cœur gisent une multitude de pratiques langagières qui sont autant de pratiques sociales matérialisées, déposées, à interpréter. L'analyse de données textuelles constitue, en ce sens, une des formes de l'analyse du social. Comme le remarque B. Lahire, « la sociologie a constamment à faire au langage, quels que soient ses méthodes et ses domaines de recherche… le langage étant présent dans toute pratique sociale.) [99, LAHIRE] », « les multiples pratiques langagières… sont indissociables de toute pratique sociale, elles sont immanentes aux liens sociaux. Les formes que prennent les relations sociales sont indissociables des pratiques langagières qui en sont une dimension constitutive. [100, LAHIRE] »

3.2.5 Les méthodes informatisées comme renouvellement des pratiques de recherche

Pour conclure, la diffusion de l'informatique et des nouvelles technologies de communication dans toutes les sphères de la société nous parait à ce point « pervasive » qu'il semble pratiquement impossible de s'en passer. Autant du côté des injonctions institutionnelles que de celui de ses pratiques quotidiennes, la recherche peut difficilement faire l'économie d'un investissement dans ces techniques. Le bouleversement conséquent de ses méthodes de travail, de ses objets de recherche, de ses problématiques demande au minimum une compréhension des principes de fonctionnement de ces outils et technologies. C'est pourquoi, outre la considération de leurs intérêts intrinsèques, la prise en compte de l'inscription nouvelle mais inéluctable de ces méthodes au sein des pratiques scientifiques réclame attention. Pourtant, leur adoption par la communauté des chercheurs en sciences sociales est encore faible en France et leur usage récent. Comme le soulignent [31, DEMAZIERE et al.] cette situation est paradoxale : les éléments langagiers constituent une

source essentielle d'analyse pour les sociologues mais sans que cette matière soit traitée autrement que manuellement, se confinant à des corpus de très petite taille.
Cette réticence (due à une méconnaissance ?) à leur égard peut constituer un frein, non seulement à l'exploration de matériaux textuels de grande ampleur et d'un type nouveau, mais aussi au renouvellement des démarches scientifiques.

Pour résumer, en reprenant les conclusions d'une enquête menée en 1987 par M.-A. Polo de Beaulieu auprès de divers chercheurs utilisant la lexicométrie, l'intérêt de ces méthodes est de permettre « l'exhaustivité (qui dépasse les capacités humaines), le brassage simultané d'une grande quantité de variables, la précision des mesures effectuées, les possibilités de comparaison de ces mesures obtenues dans les mêmes conditions, l'accessibilité et le remaniement possible des données, la caution d'objectivité fondée sur le fait que les classements sont issus des données elles-mêmes et non pas d'une construction élaborée à priori par le chercheur [55, POLO DE BEAULIEU] ».

Nous verrons, dans les paragraphes qui suivent, que malgré leurs avantages indéniables, les méthodes lexicométriques demandent à être appliquées avec certaines précautions, en raison des postulats sur lesquels elles reposent.

3.3 Limites de l'analyse de données textuelles

L'analyse textuelle, et plus largement, le recours à des techniques statistiques, bien que de plus en plus répandus, ne font pas l'unanimité, notamment en sciences sociales, comme nous l'avons vu. Ainsi, selon Paillé (dans [47, LEJEUNE et BENEL]), « la recherche qualitative [n'a] pas besoin de chiffres et ce, par conception. Adjoindre des comptages risquerait de la dénaturer, de produire des analyses quasi-qualitatives (Paillé, 2006) ». Ce rejet sans appel n'est pas sans fondement. Car malgré des qualités réelles, esquissées ci-dessus, l'emploi de l'analyse textuelle comporte des risques certains dont il faut avoir conscience et connaissance pour pouvoir les circonscrire. L'un de ces risques consiste à minorer, voire à nier, les présupposés théoriques et épistémologiques dont les outils sont une incarnation. Comme pour n'importe quelle démarche scientifique (expérimentale, qualitative, etc.), l'absence de réflexivité sur la démarche fait courir un risque à l'intégralité du processus. S'ajoute néanmoins, dans le cadre de l'utilisation d'outils informatisés, un écueil particulier, lié à la séduction conjuguée et renforcée du chiffre et de l'image. En forçant le trait, nous pourrions dire que l'apparente fiabilité accordée aux résultats provient d'une croyance à la fois dans la capacité du chiffre et de la quantification à capturer le « réel » et dans celle des graphiques à représenter ce réel sous une forme d'emblée admise – ou du moins, dont les processus mêmes d'élaboration ne sont pas systématiquement suspectés. Le statut de valeur probatoire[25] de plus en plus octroyé à ces deux artefacts dans nos sociétés participe de cette croyance. Les dénonciations de cette « illusion », des chiffres notamment, ne datent pas d'aujourd'hui. Ainsi, en 1994, Loïc Blondiaux se demandait à quoi tenait, justement, la « réussite des formes statistiques[26] ». Mais en rappeler l'existence de nos jours

[25] Voir, au sujet de l'image, la récente exposition « Images à charge », organisée au Bal, qui entendait rapprocher la construction de la preuve, notamment judiciaire, par l'image de ses conditions de production et d'exposition : http://www.le-bal.fr/fr/mh/les-expositions/forensic/. Ainsi « Comment les traces, les signes ou les symptômes d'un acte criminel peuvent-ils être découverts, compris et validés par l'image ? Comment les dispositifs de capture ou de présentation de l'image sont-ils conçus par les experts pour renforcer son caractère probatoire ? Comment l'image se construit-elle dans un discours scientifique et historique de vérité ? », *in* DUFOUR D. (dir.), *Images à Charge. La construction de la preuve par l'image*, Paris, Editions Xavier Barral - Le Bal, 2015.
[26] Pour un aperçu de la littérature sur le sujet, voir :
http://books.openedition.org/pressesmines/925?lang=fr ; et BLONDIAUX L., « Le chiffre et la croyance.

n'est pas inutile, spécialement dans le cadre d'une utilisation croissante d'outils fondés sur des méthodes statistiques, générant une profusion de graphiques. Ne pas se laisser séduire, convaincre même, par l'objectivité de ces calculs et de ces formes demande donc de s'inscrire dans un rapport distancié à la machine, de porter une attention particulière aux conditions de production des résultats, dont l'immédiateté et l'apparente simplicité concourent justement à rendre invisibles leurs processus d'élaboration.
C'est pourquoi nous allons, dans les paragraphes qui suivent, exposer quelques-uns des principes, statistiques et linguistiques, sur lesquels se fondent les outils lexicométriques et qui en font de véritables « épistémologies embarquées [31, DEMAZIERE et al.] ».

3.3.1 Pas de neutralité des outils : postulats quantitatifs

Loin d'être épistémologiquement neutre donc, l'analyse de données textuelles repose de fait sur un ensemble d'hypothèses fortes, d'un point de vue statistique et langagier. Si ces hypothèses définissent le matériau textuel à considérer et la manière dont il peut être traité, elles postulent plus fondamentalement la puissance accordée aux chiffres et à un certain type de calcul, capable de dégager une structure signifiante des matériaux textuels, eux-mêmes souvent réduits à leurs instances éminemment représentatives, les mots pleins.
Cette croyance en la capacité de l'analyse textuelle à extraire le « réel » des données textuelles provient, selon G. de Peretti de Benzécri, lequel « ne s'est pas contenté d'importer en France les méthodes d'analyse factorielle. Il était guidé par une ambition théorique et philosophique et avait, dès le départ, pour objectif d'appliquer ces techniques à l'étude de la langue : "C'est principalement en vue de l'étude des langues que nous nous sommes engagés dans l'analyse factorielle des correspondances" (Benzécri, 1981). Son objectif assez démesuré est celui de faire naître des données une structure du réel, de dégager de "la gangue des données, le pur diamant de la nature véridique". Les méthodes de quantification du texte permettraient d'obtenir une certaine neutralité du fait de l'automaticité des calculs. [85, PERETTI DE] » J. Jenny, de son côté, met en garde contre l'implicite contenu dans ce postulat benzécriste « si dominant en France qu'on ne perçoit plus ce qu'il a de relatif, de non-évident, et qu'on en viendrait à oublier les méthodes traditionnelles d'une observation quasi-expérimentale de la réalité sociale, instruite par des paradigmes explicités et des "prénotions" consciemment assumées – dans une perspective résolument constructiviste. Rappelons les termes de ce postulat : "Il existe UNE structure DU réel…" que "l'Analyse des Données" va pouvoir dévoiler au chercheur qui aura fait table rase de ses prénotions naïves. [79, JENNY] ».
Si l'on exclue cet objectif d'extraction du réel, pour le moins « démesuré », comme le souligne G. de Peretti, il n'en demeure pas moins qu'un certain nombre de postulats épistémologiques doit être accepté dans toute démarche d'analyse de données textuelles, parmi lesquels Lejeune et Benel relèvent :
- le nombre compte ;
- la représentativité détermine la validité ;
- la force des corrélations et la significativité des différences se mesurent au moyen de tests standardisés par la statistique. [47, LEJEUNE et BENEL]

Un texte serait donc réductible en une somme d'éléments, dénombrables, sur lesquels il serait possible d'appliquer des outils statistiques classiques (tableaux de fréquences, tris croisés, analyse factorielle, etc.), en vue d'en extraire, selon [45, LEJEUNE], les « propriétés objectivement observables » (répétition de termes, cooccurrences, surreprésentation ou sous-représentation de motifs linguistiques, etc.). Or, « une telle approche quantitative repose sur l'hypothèse que les phénomènes de récurrence ont une pertinence. Dès le début du vingtième siècle, ces techniques furent développées – notamment par Harold Lasswell – pour analyser la propagande dans la presse. » [31, DEMAZIERE et al.] se demandent

---

L'importation des sondages d'opinion en France ou les infortunes d'une opinion sans publics », *Politix*, 1994, vol. 7, n° 25, p. 117‑152.

d'ailleurs « quel est le statut d'opérations telles que le recensement d'occurrences, d'associations, de répétitions et de substitutions (versus, respectivement absences, oppositions, singularités, complémentarités) ». Ce primat de la récurrence est tel qu'il explique la non considération des hapax[27] de la plupart des analyses d'outils lexicométriques.
Par ailleurs, si la récurrence fait sens pour cette méthode, l'organisation interne, propre au texte et au discours, en est absente[28]. La prise en compte de la dimension syntaxique, de l'enchaînement des fragments est laissée de côté, au profit d'un dénombrement d'unités textuelles, les mots ou séquence de mots. En ce sens, le texte est considéré comme un « sac de mots », à l'intérieur duquel des spécificités lexicales émergent grâce à leur traduction quantitative.
En procédant à une réduction du matériau textuel à ses seules propriétés matérielles, quantitativement révélées, les outils lexicométriques ne peuvent pas prendre en charge ce qui dépasse le quantitatif. Comment traiter les sous-entendus ? Les références extra-textuelles ? Comment décider de la synonymie, de la synonymie ou de la modalisation qui exprime le degré d'adhésion du locuteur à son énoncé ?

3.3.2 Postulats linguistiques

La lexicométrie réduit donc le texte en un ensemble de mots pour pouvoir leur appliquer des méthodes statistiques qui en dégageront certaines régularités, sinon signifiantes du moins significatives. Cette réduction de la multiplicité à un nombre plus limité de formes normalisées s'avère nécessaire, comme nous l'avons vu, dans la mesure où elle permet de restreindre le tableau lexical sur lequel sont appliqués les calculs statistiques. N'importe quel corpus textuel constitué d'un certain nombre de formes graphiques distinctes voit son nombre total de mots augmenter par le nombre d'apparitions de ces formes au sein du corpus (nombre d'occurrences[29]). Etant donné cette profusion lexicale, l'objectif premier est de réduire le nombre de mots (formes graphiques) que l'on prendra en compte dans les analyses statistiques. « Cette réduction a pour but de faciliter et rendre plus robuste les calculs, qui seront menés par la suite, afin d'éviter de travailler sur des tables immenses et pleines de zéro. [85, PERETTI DE] »
Mais comment opérer cette réduction générale du lexique et sur quels critères la fonder ? Que retenir ? Qu'exclure du dénombrement ? En un mot, que compter ? Et quelles sont les conséquences de ce comptage ? Car tout procédé de dénombrement, de mesure modifie, par son interaction même avec l'objet mesuré, cet objet. « D'une certaine façon, mesurer c'est agir sur le sujet. [85, PERETTI DE] » Comme nous allons le voir, la réduction du lexique, en lexicométrie, implique des choix sur le langage.

3.3.2.1 Que compter ? Lemmatisation et influence de la « forme canonique » du mot

Les textes étant composés de multiples mots, ceux-ci apparaissant eux-mêmes de nombreuses fois et sous diverses formes (genre et nombre d'un nom, forme conjuguée d'un verbe par exemple), certains outils proposent de procéder à une « rationalisation » du vocabulaire. Cette rationalisation s'appelle la lemmatisation[30].

---

[27] Selon le TLFI, un hapax désigne un vocable n'ayant qu'une seule occurrence dans un corpus donné.
[28] Nous ne traitons ici que de la lexicométrie et non de la logométrie qui, au sens de Damon Mayaffre, dépasse précisément la lexicométrie, en prenant en charge « toutes les unités linguistiques jugées pertinentes du discours : mots graphiques, lemmes, cooccurrents, codes grammaticaux, enchaînements syntaxiques, etc. », MAYAFFRE D., « L'analyse du discours assistée par ordinateur », 2009.
[29] Diverses tentatives de modélisation de la gamme des fréquences d'un corpus ont été élaborées, ainsi de la loi de Zipf ou le diagramme de Pareto. Voir, à ce sujet, LEBART L. et A. SALEM, *Statistique textuelle*, *op. cit.*
[30] D'après Lebart et Salem, la lemmatisation ramène « les formes verbales à l'infinitif, les substantifs au singulier, les adjectifs au masculin singulier, les formes élidées à la forme sans élision », « elle

Le processus de lemmatisation, fonctionnalité proposée par de nombreux outils d'analyse textuelle et notamment les deux logiciels utilisés dans le cadre de cette étude, Calliope et Alceste, est l'opération linguistique qui consiste, après avoir identifié les formes graphiques[31], à les réduire à leur forme dite « canonique », autrement dit une forme débarrassée de ses formes fléchies (pluriel, conjugaison, etc.). En cela, elle correspond à une neutralisation de la flexion[32].
L'objectif de cette procédure, dans le cadre de l'analyse de données, correspond au double objectif de réduction du nombre de mots et de limitation de la perte d'information. « L'idée est de regrouper sous un même lemme[33], différents mots dont le sens est identique afin de lui donner plus de poids, mais surtout afin d'éviter de ne pas les prendre en compte du fait de la disparité des formes utilisées. [83, PERETTI] »
Or, le choix de ramener les mots à des formes normalisées, s'il permet de traiter la multiplicité des formes, peut également représenter une perte, voire distorsion, de l'information, parfois dommageable à l'interprétation des contenus. La lemmatisation opère ainsi une transformation du lexique qui est loin d'être anodine et il n'est pas certain que le regroupement de formes fléchies au sein de leur lemme conserve l'intégralité des sens. Ainsi, selon Lebart et Salem, « dans le domaine de l'étude des textes politiques, les chercheurs ont constaté que le singulier et le pluriel d'un même substantif revoient souvent à des notions différentes, parfois en opposition (cf. par exemple l'opposition dans les textes récents de *défense de la liberté* / *défense des libertés* qui renvoient à des courants politiques opposés). [82, LEBART et SALEM] » Ou encore, un verbe à l'infinitif n'indique pas le rapport au temps que les formes conjuguées peuvent exprimer.
Le choix de pratiquer (ou non) cette normalisation dépend, de fait, de plusieurs facteurs (type de corpus, finalité du projet, problématique envisagée, etc.[34]). C'est pourquoi elle rassemble aussi bien des défenseurs (la réduction de la diversité lexicale permet de mettre en avant la proximité sémantique) que des opposants (le regroupement de formes est inapproprié à l'expression de sens différents)[35].

3.3.2.2 Suprématie du mot plein sur le mot-outil

Mais le processus de lemmatisation ne consiste pas seulement à réduire la diversité lexicale. Elle s'inscrit de fait dans un paradigme « documentaire » qui postule le primat du « mot plein » sur les autres éléments du langage. La désignation même de « mots pleins », par opposition aux « mots vides » (ou mots-outils) est déjà suffisamment éloquente : le langage serait ainsi composé de mots ayant du sens et de mots ayant « seulement » une fonction syntaxique, vide de sens.

Qu'entend-on par « mots pleins » ?
Il s'agit des mots du langage ayant cette particularité de cristalliser en eux un condensé sémantique que les autres mots ne posséderaient pas. Il faut entendre par là les noms, verbes, adjectifs, adverbes (par opposition aux « mots-outils » que sont les prépositions,

---

consiste à repérer, pour en faire l'entrée, l'unité formelle commune à toutes les formes variables sous lesquelles il peut apparaître (singulier ou pluriel, masculin ou féminin pour les noms et les adjectifs ; divers modes, temps ou personne pour les verbes) », selon PETIT G., « Lemmatisation et figement lexical□: les locutions de type SV », *Cahiers de lexicologie*, 2003, vol. 82, n° 1, p. 30‑57.

[31] Une forme graphique est « une suite de caractères (lettres) entourées par des caractères délimiteurs (blancs, points, virgules…) », *in* GUÉRIN-PACE F., « La statistique textuelle. Un outil exploratoire en sciences sociales », *Population*, 1997, vol. 52, n° 4, p. 865‑887. Il s'agit le plus souvent du mot tel que rencontré dans le texte d'origine.

[32] « En morphologie, domaine de la linguistique, on nomme flexion l'ensemble des modifications subies par le signifiant des mots d'une langue flexionnelle (ou agglutinante, dans une moindre mesure) pour dénoter les traits grammaticaux voulus. » : Wikipédia, https://fr.wikipedia.org/wiki/Flexion_linguistique

[33] Le lemme correspond, de fait, à l'entrée lexicale des dictionnaires, forme canonique du terme.

[34] S'il s'agit, par exemple, de recherche documentaire, l'interrogation par lemme s'avèrera pertinente ; de même, dans le cas d'une étude générale des thématiques d'un corpus.

[35] Pour en savoir plus, voir GUÉRIN-PACE F., « La statistique textuelle. Un outil exploratoire en sciences sociales », *op. cit.*

articles, conjonctions, etc., définis principalement par leur rôle syntaxique). Ce substrat sémantique attribué aux mots pleins expliquerait donc le traitement préférentiel qui est le leur dans les outils d'analyse de données textuelles. Compter et analyser le lexique d'un corpus, cela revient (souvent) à ne compter et analyser que ses mots pleins. Certains logiciels, comme Alceste, donnent certes accès aux mots-outils, mais sans que ceux-ci n'interviennent dans l'élaboration de la classification. Ainsi, si « les types grammaticaux sont distingués dans la plupart des logiciels, certains rejettent dans le néant du non-calcul (quitte à les réintégrer en "variables supplémentaires" après calcul) tous les mots qui ne sont ni substantifs ni adjectifs ni verbes et adverbes, et qu'on appelle pour cela des "mots-outils" ou mots "vides de sens". Or, quand on connaît la richesse des informations qu'on peut extraire des usages discursifs des déterminants, pronoms personnels, prépositions et conjonctions, adversatifs, déictiques, tournures stylistiques et rhétoriques (informations de type sociologique, sur les rapports sociaux qui se construisent et s'actualisent dans les pratiques discursives et langagières), cette pseudo-neutralité méthodologique ne peut s'expliquer que par la crainte d'être contaminé par la subjectivité des chercheurs et/ou par la crainte de ne pouvoir s'acquitter d'opérations aussi complexes et subtiles – car elles le sont, effectivement. [79, JENNY] »

Comment expliquer ce primat du mot plein ?
D'abord par la profusion lexicale énoncée plus haut. En effet, les mots-outils se distinguent par leur fréquence élevée au sein des textes, contrairement aux mots pleins, bien moins fréquents. En cela, « l'approche lexicométrique peut paraître paradoxale, comme le rappellent [42, LABBE et LABBE] : "Peu de mots dépassent le seuil de 1% de fréquence relative et ce ne sont probablement pas les plus intéressants puisque, selon le vieil adage classique, la quantité d'information véhiculée par un mot est inversement proportionnelle à sa fréquence d'apparition." [85, PERETTI DE]» Ce phénomène est confirmé par l'observation de la fréquence des termes au sein de notre corpus (déclarations d'associations au JO), dans lequel les huit mots les plus fréquemment utilisés sont effectivement des mots-outils (« et », « à », « en », « pour », « qui », « entre », « dans », « ou ») qui représentent 12% du vocabulaire total. En évacuant du traitement ces termes très fréquents mais « sémantiquement faibles », les outils lexicométriques évitent que les résultats soient pollués par le poids de ces termes mais, ce faisant, interviennent profondément sur le matériau textuel. Or, « aucune de ces méthodes de compression de données (mots trop fréquents ou rares) n'est cependant neutre. Elles reviennent toujours à modifier la définition initiale du contexte et affectent les résultats. [97, HABERT, NAZARENKO et SALEM] » Il s'agira de s'en souvenir au moment d'interpréter les résultats.

Cette prédominance d'un certain type lexical proviendrait, en outre, selon, [86, PETIT], d'un impensé lexicologique, selon lequel « seules sont concernées les unités syntaxiques minimales, N-V-Adj, lesquelles constituent le lexique prototypique tant pour la conscience courante que pour celle du linguiste ». Cet impensé se voit motivé par le fait d'attribuer un identifiant linguistique à une forme graphique – identifiant normalisé qui signe l'appartenance du mot à la langue en général, et au lexique en particulier. Dans la finalité lexicologique, chaque unité ou séquence lexicale doit en effet « être répertoriée à l'inventaire de la langue et nécessitent pour ce faire une forme identifiante qui permette de les distinguer tout en manifestant leur identité sémiotique. »
Cette suprématie du mot, et du mot plein, au sein des matériaux linguistiques relève également de la pratique documentaire : le mot plein, normalisé, serait le plus à même de représenter, de manière synthétique, l'information contenu au sein des documents – et en faciliterait le repérage et la consultation. L'on parle à ce propos de « mot-clé », d'« indice » ou de « descripteur » : les mots pleins servent à décrire le contenu, et les retenir comme base du calcul statistique permettrait, précisément, d'accéder à ce contenu. Il n'est qu'à consulter un dictionnaire ou un thésaurus – ainsi que leurs corollaires actuels que sont les ontologies et les concepts du web sémantique, pour se convaincre de la pérennité de ce primat.

Ce processus de normalisation lexicale, explique, en outre, le rejet, à la fois au sein des études lexicologiques et à fortiori, des outils lexicométriques, dont sont victimes les syntagmes (nominaux, verbaux, prépositionnels, etc.). Sont ainsi exclus du traitement statistique de nombreux outils, les diverses séquences polylexicales (séquences composées de plusieurs termes[36]) car elles ne constituent pas des « unités de base ». J. Jenny s'interroge d'ailleurs sur cet abus qu'on appelle « "l'entrée lexicale", utilisée dans la plupart des logiciels d'analyses textuelles, là où on pourrait s'attendre à utiliser au moins les syntagmes nominaux. Contre les abus de l'"entrée lexicale", ne pourrait-on pas investir davantage les acquis de la sociolinguistique dans les méthodes d'analyse de corpus textuels ? » L'introduction des segments répétés au sein des outils d'analyse textuelle constitue néanmoins une tentative de prise en compte de ces syntagmes, en rétablissant une partie de la complexité et de la richesse lexicale et discursive dont l'avait privée la prise en compte du simple mot plein.

Sans vouloir aller plus loin dans l'exposé des présupposés statistiques et linguistiques des outils lexicométriques – d'autres limites existent en effet, telles que la non prise en compte des formes interrogatives ou négatives, de l'expression de la position du locuteur vis-à-vis de ses énoncés, de la tonalité des sentiments[37] (marqueurs de modalité), la difficulté à distinguer les sens métaphoriques, etc. –, nous aimerions simplement conclure en précisant que seule la connaissance des principes sous-jacents aux outils permet de choisir l'outil adapté à la finalité du projet. Ainsi, « selon les objectifs visés par la recherche, on pourra se demander si les logiciels prennent en compte les reformulations, les hésitations, les silences. Y a-t-il d'autres indices linguistiques qui traduisent la façon dont l'énoncé est pris en charge, assumé ou non par le locuteur ? Les logiciels sont-ils aptes à fournir une classification pertinente des chiffres, verbes, adjectifs, connecteurs, opérateurs ou formes de la modalisation ? [31, DEMAZIERE et al.] »

3.3.3 Constitution de corpus : une étape délicate

La constitution est une étape essentielle et délicate dans l'analyse de données textuelles dans la mesure où les résultats en dépendent. Cette complexité s'accroît avec l'accès à de nombreuses sources numériques. Le Big Data ne signifie ni facilité de traitement ni résultats immédiatement pertinents : « Big is too big » ?

[54, OLLIVIER] pointe quelques-unes des difficultés sous-jacentes aux gros corpus (dont notre propre corpus fournit un exemple frappant) :
- Dimensionnement et faisabilité de l'analyse sur corpus : sous-estimer l'importance stratégique du travail de constitution du corpus ;
- Hétérogénéité des critères et des modes d'indexation, de la structuration des données (entre plusieurs jeux de données ou pour un même jeux de données sur des périodes différentes ;
- Non homogénéité des données ;
- Possibilité de mutualiser les corpus ? (mythe ou réalité ?).

Ces difficultés sont patentes dans la perspective de l'analyse secondaire des données, dont l'objectif est la mise à disposition et la réexploitation de données. Ainsi, le questionnement

[36] Ces séquences polylexicales jouissent actuellement d'une attention particulière dans le domaine du TAL (traitement automatique des langues). Elles désignent « l'ensemble des constructions relevant du domaine linguistique de la phraséologie : collocations, idiomes, locutions, constructions à verbe support... », voir notamment CORMAN J., « Extraction d'expressions polylexicales sur corpus arboré ».
[37] Voir, à ce propos, les récents développements de l'« opinion mining » et du « sentiment analysis », qui visent à identifier et à extraire l'information subjective des textes et autres ressources matérielles, à partir de techniques issues du TAL et de la linguistique computationnelle notamment. Pour une approche approfondie, voir BOULLIER D. et A. LOHARD, *Opinion mining et Sentiment analysis*, Paris, OpenEdition Press, coll.« Sciences po / Médialab », 2012.

des méthodes de constitution de corpus (et leur documentation) devient encore plus crucial dans la mesure où les modes de collecte des données proviennent de personnes autres que le chercheur, orientées vers des finalités différentes. [31, DEMAZIERE et al.] se demandent d'ailleurs « comment [le sociologue] peut-il se représenter « son » corpus lorsque celui-ci provient de sources multiples et qu'il est exploitable à l'infini ? Quels sens peut-on donner à la trace des données textuelles ainsi préservées ? ». Le lien de proximité aux données qui caractérisait le sociologue tend à se distendre avec l'accès à de multiples sources, réutilisables.
Si les outils d'analyse de données textuelles paraissent adaptés au traitement des grands corpus du futur (corpus cumulatifs, hétérogènes, constitués à partir de techniques différentes et issus de sources variées), c'est sans compter les multiples questions juridiques, épistémologiques, méthodologiques et pratiques que cela soulève !

Nous aimerions ajouter néanmoins que profusion de données n'est pas profusion de corpus. [21, RUIZ] précise ainsi que « la massification de la documentation consultable ne signifie pas forcément massification de la documentation intéressante. En ce sens, l'échelle est différente, mais la question est ancienne : apprendre à sélectionner les documents et à combler les vides de l'information est depuis longtemps au cœur du métier d'historien. Ici, c'est à un retour aux sources de la méthode critique historique (de l'école méthodique à celle des Annales) qu'il est nécessaire d'appeler. » L'avènement de l'accès démultiplié aux sources d'information doit donc être relativisé.

3.3.4 Risques de dérapage interprétatif en puissance

Méconnaître l'influence des principes de fonctionnement des outils sur les résultats – ou plutôt, ne pas y voir le reflet même de leurs fondements épistémologiques – constitue donc un enjeu majeur pour la fiabilité des résultats produits au moyen de ces outils. Le risque de dérapage interprétatif rode toujours. Ainsi, ce risque « se pose en analyse textuelle comme dans d'autres types d'études mais les données textuelles se prêtent sans doute davantage aux "projections" interprétatives. Dans tous les cas de figure, l'analyse produite est inséparable des conditions de production des données, c'est-à-dire l'ensemble des choix opérés par l'analyste à chaque étape de l'étude (constitution du corpus, choix du logiciel et de ses paramètres). [28, BRUGIDOU et al.] » Et [30, DELAVIGNE] de noter, à propos de la méthode employée par Alceste, que « la construction de classes peut laisser croire que le logiciel livre une « vérité intrinsèque » sur le corpus mais il s'avère que, dès lors que l'on change quelques paramètres (modification des variables par exemple), ces classes peuvent changer ». Les résultats ne sont donc pas une description, un reflet du réel (d'ailleurs, que serait ce réel ?) mais bel et bien le fruit d'une construction.
Ce risque d'acceptation d'une vérité intrinsèque des résultats se double de celui de « raconter une histoire », fondée sur la sélection de certains résultats parmi l'ensemble mis à disposition, pouvant conforter les hypothèses de recherche. Ce biais, présent dans toute démarche scientifique, l'est d'autant plus avec les outils d'analyse textuelle que leurs résultats paraissent « objectifs ».

3.3.5 Risque de soumission de la démarche scientifique

Cette méconnaissance des principes de fonctionnement des outils présente un autre danger, plus large et plus fondamental, qui concerne la démarche scientifique elle-même : si cette croyance en l'objectivité des techniques et outils informatisés n'était pas questionnée, elle représenterait une soumission de la réflexivité scientifique à la tentation techniciste. « Il est illusoire de penser que l'intuition et la subjectivité sont écartées par l'usage de procédures informatisées, sauf à prétendre que le métier de sociologue consiste pour une bonne part à déléguer les opérations de recherche à un pilote automatique. […] Aussi les usages des logiciels d'analyse textuelle doivent-ils rompre radicalement avec toute tentation techniciste qui y verrait un moyen décisif pour réduire la place de l'intuition du chercheur dans l'activité analytique et pour éradiquer le bricolage qui serait irrémédiablement associé aux traitements

manuels. Car le traitement informatisé ne situe pas magiquement le chercheur du côté de la rigueur. Au contraire, il l'expose aussi à une dérive vers des pratiques troubles, opaques pour ces pairs et même, pourquoi pas, imperceptibles par lui-même si elles se coulent dans les protocoles programmés et sont dictées par les contraintes de l'outil. »

3.4 Intérêts de l'analyse de données textuelles

Après montré quelques-uns des pièges et difficultés émaillant toute démarche d'analyse de données, nous aimerions désormais insister sur les garde-fous capables de les déjouer. La mise en avant de ces difficultés n'avait d'autre fonction que de permettre de les cerner pour mieux les dépasser. Car les intérêts de cette démarche, indéniables, résident dans la mise à distance de ses limites intrinsèques. Autrement dit, seule la connaissance des risques et des biais encourus rend possible leur circonscription et peut constituer l'analyse textuelle en une véritable aide à la démarche scientifique. Nous essaierons dans les paragraphes qui suivent de répondre à quelques-unes des objections à la lexicométrie soulevées dans la partie précédente.

Notre réponse aux limites de l'analyse de données textuelles se décline en deux temps :
1) L'analyse de données doit, tout d'abord, être considérée comme relevant d'un processus scientifique global, ie un processus réflexif, quanti-qualitatif, itératif, à finalité exploratoire.
2) Lors de sa mise en pratique, l'analyse de données doit intégrer certaines précautions : élaborer le corpus en tenant compte de certains principes ; en fonction de la problématique, choisir des outils possédant un processus de lemmatisation amélioré (grâce à l'introduction de connaissances linguistiques plus sophistiquées) ; limiter les dérapages interprétatifs par un retour aux textes, une analyse de ce qui est étonnant ou rejeté des résultats et une documentation de toutes les étapes de la démarche ; accompagner les résultats graphiques d'un discours critique ; prendre en considération le facteur temps dans tout projet d'analyse de données, notamment dans l'assimilation des compétences requises.

3.4.1 L'analyse de données comme partie d'un processus scientifique global

3.4.1.1 L'analyse de données textuelles : une explicitation du processus réflexif

Malgré des différences d'échelle, d'ampleur et de méthodes, aucune différence de nature ne sépare démarche manuelle et démarche intégrant des outils informatisés. L'apparente facilité des outils informatisés ne le cède en rien, en termes de réflexivité, à une démarche manuelle. La manipulation de gros corpus numériques, la médiation de la machine n'oblitèrent pas le recul et le questionnement incessants, propres à la posture scientifique.
Comme le notent [31, DEMAZIERE et al.], « les usages des logiciels d'analyse textuelle doivent être inséparables d'une réflexivité organisant des va-et-vient entre le traitement des matériaux et le travail interprétatif, entre les analyses empiriques et les conceptualisations. La sophistication des instruments ne doit jamais faire perdre de vue les hypothèses, les questionnements et les problématiques, qui doivent guider tout cheminement de recherche. Car les traitements informatisées, les plus rudimentaires comme les plus sophistiqués, ne sont jamais neutres, pas plus que les simples calculs de fréquence ou les classifications hiérarchiques.[38] » M. Reinert, lui-même, concepteur de la méthode Alceste, souligne qu'il faut « mettre en garde l'utilisateur d'une technique, lui rappeler que celle-ci n'est pas neutre, qu'elle ne dispense pas de s'interroger sur le sens d'une démarche même si la forme que prend cette interrogation peut paraître naïve et hors du champ de la science. Si le sens reste

[38] L'explicitation des opérations réalisées dans le cadre du projet « Mémoire » constitue précisément l'objectif des chapitres suivants : « Méthodologie » et « Résultats ».

fondamentalement incommunicable, lié à la liberté d'être de chacun, une technique qui n'aurait pas de sens, n'est qu'une technique d'aliénation.[39] [61, REINERT] »
Le questionnement met donc à distance les pièges possibles de l'analyse de données textuelles par l'explicitation, à chaque étape du processus, des opérations effectuées sur les matériaux : collecte, structuration, codification, réduction, exploitation, interprétation. L'adéquation aux problématiques de recherche doit ainsi devenir la pierre de touche des multiples choix opérés.
Les biais potentiels de l'analyse textuelle, peut-être moins connus que ceux des démarches plus traditionnelles, offrent, de façon paradoxale, un avantage inestimable : ne pouvant en faire l'économie, leur prise en compte permet de rendre visible le processus de recherche. Si l'utilisation des outils ne saurait en effet se substituer à la réflexivité fondamentale propre à la démarche scientifique, elle permet néanmoins d'en expliciter le processus. « En effet, l'objectivation des procédures de traitement dans des lignes de commande permettant de piloter des algorithmes, et la formalisation des résultats obtenus dans des graphes, tableaux, et autres schémas, permettent de décomposer plus aisément que dans des analyses manuelles ou artisanales les opérations effectuées par le sociologue. [31, DEMAZIERE et al.]»

3.4.1.2 Une méthode quanti-qualitative

Par ailleurs, toute démarche intégrant des outils lexicométriques se révèle proprement quanti-qualitative. La fausse opposition entre deux types de démarches, souvent dénoncée[40], voudrait distinguer d'un côté, des méthodes purement quantitatives (traitement statistique, dénombrement, quantification), et de l'autre, une démarche purement qualitative, d'interrogation du social, d'élaboration de problématiques, étudiant des données qualitatives. Or, l'usage de méthodes statistiques au sein d'une démarche qualitative n'est pas la simple juxtaposition des deux démarches, la coexistence de deux méthodes hermétiques. Elles constituent bel et bien une seule démarche, globalement quanti-qualitative où le qualitatif « in-forme » sans cesse le quantitatif qui, lui-même, transforme le qualitatif en le traduisant dans un autre langage. Nous avons montré plus haut à quel point les principes statistiques reposaient sur des constructions, des hypothèses non quantitatives (telle la signification prédominante accordée à la fréquence). L'intégralité du processus d'analyse textuelle est donc traversée de qualitatif et de quantitatif, avant, pendant et après le traitement proprement dit : choix de la source, constitution du corpus, choix du type d'outils et à travers eux, du type de calculs effectués, choix des unités de base (lemmatisation ou non du vocabulaire), choix des paramètres, analyse et interprétation des résultats, etc.
De manière plus large, peut-il même exister des démarches purement qualitatives ou purement quantitatives ? Qu'est-ce donc que l'annotation manuelle d'un corpus, si ce n'est sa réduction à un certain nombre d'étiquettes[41], elles-mêmes transformées en code (donc en valeur numérique), en facilitant l'interprétation ? Certains traits (langagiers, sociologiques, etc.) sont-ils majoritaires ? Ne voit-on pas déjà poindre du quantitatif, une tentation de

[39] Il est à ce propos intéressant de remarquer que M. Reinert ne désigne jamais Alceste par le terme « logiciel » mais toujours par celui de « méthode ».
[40] Voir, à ce sujet, LEJEUNE C. et A. BÉNEL, « Lexicométrie pour l'analyse qualitative. Pourquoi et comment résoudre le paradoxe? », 2012., MAYAFFRE D., « «□Ça suffit comme ça!□». La fausse opposition quantitatif/qualitatif à l'épreuve du discours sarkozyste », *Corela. Cognition, représentation, langage*, 2014, n° HS-15., ou JENNY J., « "Quanti / Quali", Distinction fallacieuse et stérile□! », Villetaneuse, 2004.
[41] Comme le montrent LAROSE F., T. KARSENTI, et V. GRENON, « Regards sur diverses approches de traitement des données textuelles. Les outils, leurs fondements et l'épistémologie de leurs usages », *Formation et profession*, 2000, n° 6/2, p. 5‑12 : « Si vous administrez un questionnaire à 100 personnes et celui-ci inclut une rubrique d'identification qui, à son tour, offre au sujet la possibilité de se classer dans la catégorie « homme » ou « femme » de la variable « sexe », vous aurez créé deux types de données qualitatives (deux catégories) caractérisant une variable tout aussi qualitative. Que fait-on généralement de ces informations ? On leur appose une valeur numérique (un code) et on les additionne en attendant de s'en servir en tant que critère de distinction ou variable de contraste. »

dénombrement et de mesure qui permet la comparaison de données qualitatives ? De l'autre côté, qu'est-ce que le choix de variables, dans une étude quantitative, si ce n'est l'introduction du qualitatif ? Ce choix ne repose-t-il pas sur une hypothèse ? Que signifie la variable « date » si ce n'est une représentation quantitative d'un phénomène qualitatif ? Que sont les niveaux de mesure quantitatifs, si ce n'est une représentation de la variété du social ? Ainsi, selon J. Jenny, « non seulement ces deux types de méthodes sont complémentaires mais encore il ne devrait même pas y avoir lieu de les distinguer car en fait, si l'on y regarde de plus près, on s'aperçoit qu'aucune des deux ne peut fonctionner sans intégrer des éléments substantiels de l'autre. Peut-on dire [de l'analyse textuelle] qu'il s'agit d'une intrusion, voire d'une contamination, des mathématiques dans le camp du langage, ou l'inverse ? Ou n'est-ce qu'une métaphore ? Dans toute expression langagière, les dimensions dites quantitative et qualitative sont indissociables, intrinsèquement associées l'une à l'autre. Il ne saurait y avoir, dans toute information, de dimension quantitative sans dimension qualitative intrinsèquement associée et réciproquement. Toute recherche en sciences sociales comporte nécessairement une part de "matériaux textuels" à analyser – au point que la distinction entre le "qualitatif" et le "quantitatif" ne saurait être au mieux qu'une distinction de phases, de moments dans la recherche, et au pire qu'une mystification. Mystification destinée peut-être à masquer les méconnaissances respectives de la réalité discursive chez les "quantitativistes" d'une part, et de la réalité numérique chez les "qualitativistes" d'autre part. [79, JENNY] »

3.4.1.3 Une démarche itérative et exploratoire, au service du raisonnement scientifique

Réflexive et quanti-qualitative, la démarche scientifique qui fait appel à l'analyse de données textuelles est aussi, nécessairement, itérative. Si chaque étape doit être questionnée, si son adéquation à la finalité doit être constamment vérifiée, alors le cheminement menant de l'une à l'autre ne saurait être linéaire. Progressif, fait d'allers-retours, de corrections, de bifurcations, ce cheminement permet au projet de s'adapter, à chaque étape, aux problématiques. Son périmètre évolue sans cesse, en fonction des obstacles rencontrés, des choix à expliciter et ce, dans un double mouvement : adaptation du projet aux hypothèses initiales mais aussi évolution des hypothèses en fonction des différentes interprétations possibles des résultats.

En cela, les logiciels d'analyse de données textuelles servent le raisonnement sociologique, lui permettant de reformuler, de manière progressive, ses interrogations. « La réflexion sociologique est bien le guide qui oriente le pilotage des commandes adressées au logiciel, de sorte que le recours à cet instrument […] permette une traduction de l'imagination sociologique en propositions précises et opératoires. La posture, réflexive, propre à la démarche scientifique ne réduit pas les logiciels « à des instruments d'objectivation et d'administration de la preuve dont les résultats s'imposeraient au chercheur, mais une posture qui les considère comme des ressources mobilisables, parmi d'autres, pour nourrir des interrogations sociologiques, tester des lectures interprétatives, éprouver des significations provisoires ». De même, Benzécri note, à propos de l'usage de méthodes statistiques appliquées aux données textuelles, que « parmi toutes les idées à priori, souvent contradictoires, que chaque problème suscite en si grand nombre, un choix opportun s'opère : bien plus, l'idée qui à posteriori, après examen statistique des données, semble avoir été à priori fort naturelle ne se serait pas toujours présentée d'elle-même à l'esprit.[42] »

Loin donc de proposer des réponses définitives à des questions figées, le recours aux outils d'analyse textuelle, par leur exploration complexe et différenciée du matériau textuel, constitue essentiellement une aide à la reformulation de questions. Ni descriptive, ni probatoire, cette démarche est, selon nous, profondément exploratoire.

[42] Cité par BEAUDOUIN V., « Statistique textuelle ☐: une approche empirique du sens à base d'analyse distributionnelle », *op. cit.*

3.4.2 Les précautions à adopter dans la mise en pratique de l'analyse de données textuelles

Nous aimerions désormais lister un certain nombre de garde-fous, permettant de limiter les mauvais usages des logiciels lexicométriques et les dérapages interprétatifs. Pour que l'analyse de données textuelles devienne réellement une source d'inspiration à la démarche scientifique, celle-ci doit intégrer certaines précautions dans sa mise en pratique de l'outil informatisé. La précaution fondamentale, déjà évoquée, consiste à ne jamais séparer les résultats de leurs conditions de production, c'est-à-dire de « l'ensemble des choix opérés par l'analyste à chaque étape de l'étude (constitution du corpus, choix du logiciel et de ses paramètres). [28, BRUGIDOU et al.] »

3.4.2.1 Constitution du corpus, des principes à respecter

Qu'est-ce qu'un corpus textuel ?
Une des premières étapes consiste à s'interroger sur le mode d'élaboration du corpus de données. Si le corpus de données choisi conditionne les résultats[43], une attention particulière doit alors s'attacher à son élaboration. Car, comme le rappelle [94, DALBERA], « l'analyse ne vaut que ce que vaut le corpus ».
Pourtant, nous ne donnerons pas ici de définition précise du corpus[44], dans la mesure où il existe une variété de corpus, dont chacun se détermine en fonction de la problématique de recherche dans laquelle il s'inscrit, du champ disciplinaire concerné, du type de documents rassemblés, de la finalité recherchée, etc. Pour reprendre l'expression de P. Charaudeau, « Dis-moi quel est ton corpus, je te dirai quelle est ta problématique / Dis-moi quelle est ta problématique, je te dirai quel est ton corpus [93, CHARAUDEAU] ».
A défaut de fournir une définition précise et opératoire du corpus, nous tenterons néanmoins d'en donner quelques éléments de caractérisation, permettant de mieux comprendre comment un corpus peut être élaboré, notamment en sociologie

Corpus et sciences sociales
Précisons néanmoins que, contrairement à la conception du corpus en linguistique[45], le corpus tel qu'utilisé en sciences sociales (sociologie et sciences politiques notamment) renvoie intégralement, selon Damon Mayaffre, au « monde réel ». En cela, il est tout entier

---

[43] Ainsi, selon J.-P. Dalbéra, « Le type de données sélectionnées n'est jamais innocent et traduit une préoccupation sous-jacente. « Pour prendre un exemple simple, les corpus rassemblés par les dialectologues chargés d'enquêter, dans le cadre de la même entreprise, en obéissant aux mêmes consignes, afin de réaliser les Atlas Linguistiques de la France par régions ne sont pas complètement analogues. Certains dialectologues se sont souciés, lors des enquêtes, de noter les données négatives, tandis que les autres ne l'ont pas fait. Cela ne peut manquer d'avoir un impact sur les interprétations ultérieures. », *in* DALBERA J.-P., « Le corpus entre données, analyse et théorie », *Corpus [En ligne]*, 2002, n° 1.

[44] Dison simplement que, selon G. Ollivier, le corpus sociologique est une collection de textes rassemblés en vue de répondre à une problématique sociologique relevant de l'ancrage socio-historique des textes. De même, pour J. Jenny, « en sociologie et autres disciplines de sciences sociales, on peut se permettre de définir comme corpus tout ensemble de textes (d'origine écrite ou verbale, publique ou privée, mais rarement exhaustifs ni représentatifs) qu'on rassemble au cours des investigations de terrain (le terrain pouvant se situer dans la rue, le métro, à la radio, à la télévision, au bistro, au boulot, comme dans les chansons, les proverbes et autres aphorismes, dans les journaux, les livres, dans les textes de loi, [au *Journal officiel*], les circulaires administratifs, sur les affiches publicitaires murales, etc.), en fonction des objets de recherche, extra-linguistiques, (par exemple, controverses, débats, conversations, mouvements sociaux, pratiques et représentations sociales, concernant tel ou tel problème ou "fait de société", domaine d'action ou d'activité, rapport social (de classe, de genre, de génération, d'ethno-culture, etc.), évolution en cours, résistances au changement, etc. ».

[45] Voir, par exemple, la définition de John Sinclair (1996), citée par HABERT B., A. NAZARENKO, et A. SALEM, *Les linguistiques de corpus*, Paris, Armand Colin, 1997., « Un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques explicites pour servir d'échantillon du langage ».

« référentiel [102, MAYAFFRE] ». Loin d'être un échantillon représentatif du langage, la matière textuelle[46] dont le corpus est constitué, en sciences sociales, « ne vaut que » pour sa valeur extra-linguistique. Le rapport entre langue et société est d'emblée posé. Le corpus ne sert pas de support pour étudier un état de la langue mais pour ce qu'il dit (ou dirait) d'une réalité « extérieure », sociale. Il s'agit d'appréhender les phénomènes sociaux à travers les réalisations langagières des individus. En cela, il ne s'apparente pas à un échantillon (au sens statistique du terme) mais davantage à un sondage, au sens d'une technique d'exploration d'un milieu, en vue d'en extraire des composants auxquels l'interprétation donnera sens.
Le lien que la sociologie entretient au langage ne signifie pas, par contre, un rapport d'extériorité dans la mesure où le langage lui-même est un phénomène social. Si elle n'étudie pas les productions langagières en tant qu'échantillon du langage, la sociologie s'appuie sur ces productions comme autant de signes de relations sociales. Comme le souligne B. Lahire, « dès lors que l'on s'intéresse à la construction des structures cognitives, on se rend compte que la conscience de tout être social se forme et prend existence dans et par le(s) langage(s), à travers les multiples relations qu'ils nouent avec le monde et avec autrui… Il faut donc éviter l'emploi de toutes les formules qui donnent à penser que l'activité langagière n'est que l'expression de quelque chose de pré-social qui se serait formé dans la conscience hors de tous signes, expression qui serait une sorte de publicisation d'une activité intérieure, privée, intime… Les structures mentales, cognitives caractéristiques d'un être social particulier sont le produit, à un moment donné, des formes de relations sociales dans lesquelles il a été et est pris et qu'il contribue à maintenir ou à modifier. [99, LAHIRE] »

Corpus et représentativité

Pourquoi utiliser un corpus ? Aucune analyse, aussi ambitieuse soit-elle, ne saurait embrasser l'exhaustivité, de même qu'aucune exhaustivité ne saurait être atteinte. D'ailleurs que signifierait pareille exhaustivité ? Nombreuses sont les études pointant l'impossibilité d'une telle caractéristique. Si l'exhaustivité est illusoire, peut-on du moins prétendre à la représentativité ? Or, le caractère représentatif du corpus est lui-même problématique[47]. « Le corpus équilibré est sans doute celui qui a "de tout un peu", mais encore faudrait-il savoir ce qu'est "tout" [104, PERY-WOODLEY] » et, nous ajoutons, ce qu'est « un peu ».
Le corpus ne peut donc qu'être constitué à partir de choix en vue de répondre aux besoins d'une analyse : cela signifie élaborer des « données ordonnées, cohérentes et homogènes, et aussi représentatives que possible au regard de l'objectif en fonction duquel elles ont été constituées [105, PREVOST] ».
Le caractère nécessairement circonscrit des données à collecter et à traiter conduit à faire l'hypothèse (le pari, selon [94, DALBERA]) qu'une analyse limitée à un ensemble de textes pourrait être de nature à rendre compte de certains phénomènes sociaux. Mais s'il en rend compte, ce n'est que d'une manière exploratoire, temporaire, qui appelle d'autres questionnements, d'autres analyses, d'autres corpus. Le corpus, dans ce cadre, ne saurait être autre qu'un « prétexte », un point de départ demandant à être confronté d'autres explorations, d'autres recompositions des mêmes données (élaboration de sous-corpus) mais aussi à être comparé à d'autres corpus, à d'autres textes, à d'autres recueils de données. Aide temporaire au parcours interprétatif, le corpus ne peut donc jamais être clos ni définitif. Ainsi, selon P. Charaudeau, il s'agit de « mettre en relation les résultats d'une analyse descriptive avec ceux d'autres analyses : ceux d'autres corpus connexes (confrontation des articles de différents journaux pour en interpréter les ressemblances et différences) ; ceux de corpus de textes d'un même domaine mais de situations différentes (confrontation des écrits journalistiques de différentes époques) ; ceux, enfin, des analyses proposées par d'autres disciplines sur le même domaine discursif (philosophie, histoire, sociologie,

[46] Les données textuelles réunies au sein d'un corpus peuvent être soit collectées par le sociologue (corpus d'entretiens), soit provenir d'autres sources (comme les déclarations du JO dans notre cas).
[47] Voir notamment HABERT B., « Des corpus représentatifs□: de quoi, pour quoi, comment□? », M. BILGER (dir.), *Linguistique sur corpus. Études et réflexions*, Perpignan, Presses Universitaires de Perpignan, 2000, p. 11‑58.

psychologie sociale), sur, par exemple, le domaine politique. C'est pourquoi aucun corpus n'est jamais définitivement fermé, sa clôture ne pouvant être que le fait d'une décision provisoire à des fins opératoires. Ce sont les paramètres de "contrastivité" (externe et interne) du corpus qui par le jeu des ressemblances/différences font sens. »

Eléments à considérer lors de la constitution de corpus
Imparfaitement représentatif et temporaire, le corpus peut, en fonction des problématiques qui lui sont assignées, être composé selon diverses variables. P. Charaudeau distingue ainsi variables externes et internes, qui permettent de sélectionner les matériaux textuels sur lesquels l'analyse s'appuiera :
- **Variables externes** : comparaison temporelle de textes appartenant à différentes époques ; comparaison spatiale et culturelle de textes de différentes aires géographiques ; comparaison de différents types de discours sous l'angle d'une thématique, comme la stratégie de persuasion à l'œuvre dans les discours politique et publicitaire, etc. ;
- **Variables internes** : étude de textes provenant de la même source, du même auteur, du même support, du même contexte situationnel (campagnes électorales).

Pour conclure, le corpus doit prendre en compte, lors de son élaboration, les éléments suivants :
- La question de sa représentativité (variété de textes, d'auteurs, de sources ; texte intégral ou échantillon ; etc.) et de son homogénéité (sur le plan chronologique, narratif, des conditions d'énonciation, etc.), en fonction de la problématique visée[48] ;
- Les résultats (provisoires) de l'analyse conditionnent pour partie les questions et configurent l'exploration de futures données ;
- La question de la réutilisabilité du corpus[49], de son annotation et documentation doit être envisagée, ainsi que celle de son évolutivité[50] ;
- La disponibilité des données (documents numérisés, données structurées, nettoyées, etc.)
- La taille du corpus : le corpus doit être suffisamment volumineux pour respecter les postulats statistiques.

Bien sûr, comme le souligne E. Marshman, « certains de ces critères sont difficiles à équilibrer entre eux et représentent des difficultés dans la construction du corpus [101, MARSHMAN] ».

3.4.2.2 Le problème de la lemmatisation dépassé ?

Nous avons vu plus haut que le processus de lemmatisation pouvait comporter un certain nombre de risques. Remarquons tout d'abord, que ce problème, loin d'être nouveau, a été soulevé dès l'origine de l'analyse de données textuelles et que les différents outils de

---

[48] Pour une présentation détaillée de ces éléments (canal de communication, sphère de communication, taille du texte, niveau de complexité de l'encodage linguistique et des thématiques, mode de structuration du texte, fonction de communication dominante, statut des acteurs, niveau lexico-sémantique, niveau de construction syntaxique et textuelle, niveau extra-linguistique), voir BRUGIDOU M., C. ESCOFFIER, H. FOLCH, S. LAHLOU, D. LE ROUX, P. MORIN-ANDRÉANI, et G. PIAT, « Les facteurs de choix et d'utilisation de logiciels d'Analyse de Données Textuelles », *Cinquièmes Journées Internationales d'Analyse Statistique des Données Textuelles - Lausanne*, 2000.
[49] Cette réutilisabilité du corpus en vue d'analyse secondaire est une question complexe car elle nécessite qu'un corpus soit « à géométrie variable, capable d'être adapté à des approches et à des objectifs divers » et doit, pour cela, être le plus diversifié et le plus documenté possible, selon PÉRY-WOODLEY M.-P., « Quels corpus pour quels traitements automatiques□? », *op. cit.*
[50] A noter que selon Lejeune et Benel, « peu de systèmes sont cependant conçus pour gérer des corpus en cours de maturation. La plupart du temps, tous les calculs doivent être reconduits à chaque modification. Pour implémenter de manière efficace des calculs lexicométriques, il est nécessaire de gérer finement l'impact des mises à jour du corpus : lors de l'ajout, du retrait ou de la modification d'un texte, il s'agit de conserver les résultats intermédiaires non impactés et de recalculer le reste. », *in* LEJEUNE C. et A. BÉNEL, « Lexicométrie pour l'analyse qualitative. Pourquoi et comment résoudre le paradoxe? », *op. cit.*

lexicométrie proposent souvent des analyses avec ou sans lemmatisation du vocabulaire. Par ailleurs, selon certains chercheurs (notamment D. Mayaffre), le problème de l'identification et de la normalisation de l'unité textuelle de base en lexicométrie se verrait dépassé grâce aux progrès des logiciels de lemmatisation, dont la qualité d'étiquetage morpho-syntaxique s'est accrue depuis l'apparition des premiers logiciels d'analyse de données textuelles. Il est donc désormais possible de procéder à une lemmatisation plus fine du lexique du corpus, qui tient compte non seulement de catégories morpho-syntaxiques plus détaillées mais également d'informations d'ordre sémantique. Par l'introduction d'une dimension linguistique approfondie (syntaxique et sémantique) au sein des statistiques, l'étude des matériaux textuels bénéficie d'informations sur la nature même des composants du corpus, plus fine que le simple genre des substantifs.

Ces technologies rendent alors possible un traitement approfondi des textes, qui reposent sur des méthodes issues notamment de l'ingénierie des langues. Ainsi, un outil tel que Cordial[51] dispose d'un étiqueteur morpho-syntaxique avec lemmatiseur, qui fournit le lemme, le type grammatical, le nombre d'ambiguïtés, le groupe grammatical d'appartenance, la fonction grammaticale du groupe et des informations sémantiques (résultat de la désambiguïsation). Il propose en outre de retrouver les collocations sur plus d'1,2 milliard de mots, l'extraction de phrases de corpus, une analyse logique visualisable avec désambiguïsation sémantique, une liste des occurrences lemmatisées ou non, la recherche de mots-clés, de syntagmes-clés, de phrases-clés, de concepts-clés. En schématisant, nous dirions que l'analyse textuelle sort de l'appréhension quantitative du mot pour aborder la nature qualitative de son contexte. Selon E. Brunet[52], « qu'il s'agisse des fonctions dans la phrase, des parties du discours, ou des temps, des personnes ou des modes verbaux, on explore ici ces perspectives plus syntaxiques que lexicales [69, BRUNET] ».

Selon D. Mayaffre, le développement de ces techniques perfectionnées a permis le passage d'une lexicométrie à une véritable logométrie, une analyse du discours qui dépasse l'analyse lexicale mais sans la renier. « Mais que l'on ne s'y trompe pas cependant : le propos n'est pas de renoncer au traitement lexicométrique sur textes bruts, il est de compléter ce traitement par une analyse complémentaire sur textes lemmatisés. Mieux : cet article voudrait rappeler que le traitement des textes lemmatisés qui ouvre la voie à des analyses grammaticales ou syntaxiques nous semble indispensable, mais qu'il ne peut se faire qu'à condition de garder accès au texte réel, natif, brut, que le locuteur / scripteur a effectivement émis. » En un mot, la logométrie, pour D. Mayaffre, est « un ensemble de traitements documentaires et statistiques du texte qui ne s'interdit rien pour tout s'autoriser ; qui dépasse le traitement des formes graphiques sans les exclure ou les oublier ; qui analyse les lemmes ou les structures grammaticales sans délaisser le texte natif auquel nous sommes toujours renvoyés. C'est finalement un traitement automatique global du texte dans toutes ses dimensions : graphiques, lemmatisées, grammaticalisées. [49, MAYAFFRE] »

Cet entrelacement des statistiques et de la linguistique, des nombres et des mots représente, à nos yeux, une piste intéressante pour l'analyse de données textuelles.

---

[51] Le logiciel Cordial (http://www.cordial.fr) est développé par la société Synapse (licence payante). Autre lemmatiseur, TreeTagger permet d'annoter un texte avec des informations sur les parties du discours et des informations de lemmatisation. Il a été développé par Helmut Schmid dans le cadre du projet « TC » au sein du ICLUS (Institute for Computational Linguistics of the University of Stuttgart). TreeTagger peut également être utilisé comme un « chunker » pour l'anglais, l'allemand et le français (étiquetage des parties du discours, délimitation des groupes syntaxiques, étiquetage des groupes). TreeTagger est gratuit pour des usages non commerciaux (http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger).

[52] Il n'est pas anodin qu'E. Brunet soit linguiste de formation et de profession. Concepteur d'Hyperbase, dont une version web vient de voir le jour : http://hyperbase.unice.fr. Ce logiciel est compatible avec les lemmatiseurs Cordial et TreeTagger.

3.4.2.3 Voir ce qui est surprenant ou écarté de l'analyse, retour aux textes et documentation du processus

Afin de limiter les possibles biais introduits par la double puissance évocatrice des chiffres et des images, évoquée plus haut, il est possible de prendre certaines précautions, qui seront autant de gages de validité de l'analyse.

Si l'analyse de données textuelles permet de dresser une représentation synthétique des données, elle ne le fait qu'à titre d'indices pour l'interprétation. Outre les éléments mis en avant, d'autres éléments, surprenants, doivent alerter l'analyste. « Pour des corpus de grande taille, l'approche lexicale présente l'avantage de réduire considérablement le volume d'information à lire et à analyser, mais le calcul des propriétés statistiques du texte (richesse lexicale, indices de spécificité, segments répétés, associations...) offre surtout la possibilité de différentes lectures assistées (découvertes de résultats statistiques surprenants... donc nouvelles interrogations...). [33, FALLERY et RODHAIN] » Ou encore, comme le remarque B. Pincemin, lors d'un échange sur la liste de diffusion de TXM, « l'intérêt du calcul statistique, c'est [aussi] d'apporter des éléments nouveaux, inattendus – et il faut donc méthodologiquement bien garder la porte ouverte pour ceux-ci. »

Il peut être judicieux, par ailleurs, de repérer ce qui ne fait pas sens, d'un point de vue statistique, ie voir ce qui a pu être écarté de l'analyse (en raison de fréquences trop faibles, d'un manque de spécificité, etc.). La consultation des mots non analysés dans Alceste ou des mots bruyants[53] dans Calliope ou encore des hapax peut ainsi se révéler pertinente pour l'analyse. Tout comme la fréquence et la répétition, l'absence et la rareté peuvent constituer des signes d'un phénomène à étudier. Car « l'importance et l'impact d'un mot (c'est-à-dire l'effet de sens) ne sont pas nécessairement liés à sa récurrence : l'expression « fracture sociale » employée par Jacques Chirac lors de la campagne présidentielle de 1995, les mots « karcher » et « racaille » employés par Nicolas Sarkozy, ne furent employés qu'une seule fois par leurs auteurs, et l'effet qu'ils ont produit tient évidemment au sens qu'ils avaient dans leur contexte situationnel et interdiscursif, mais aussi au fait qu'ils ont été relayés par les médias, donc par d'autres corpus. L'impact des mots n'obéit pas nécessairement à un critère quantitatif. » Si la rareté est particulièrement importante en analyse de discours, elle peut l'être également pour d'autres champs disciplinaires. Le langage étant une réalité sociale, le fait de sur-employer ou de sous-employer un terme ne manque pas de signification. Dans notre corpus de déclarations au JO, le terme « discrimination » a été marqué comme « bruyant », en 1984, par Calliope. Que nous apprend ce manque de fréquence ? Dans la mesure où le projet dans lequel nous nous inscrivons entend justement évaluer le rôle que peuvent jouer les dites « revendications particularistes » au sein de la thématique mémorielle, ce terme peut constituer un indice intéressant pour l'analyse. « Au fond, dans le vocabulaire de la lexicométrie, les mots rares, les spécificités et les segments répétés se révèlent de bons candidats marqueurs. Tous les hapax, toutes les spécificités et tous les segments répétés ne deviendront évidemment pas des marqueurs. Mais leur identification peut, en première analyse, suggérer quelques pistes heuristiques à l'analyste. Envisager les mesures lexicométriques selon leur apport heuristique n'est pas nouveau. C'est précisément l'inscription épistémologique que Max Reinert donne à ce type d'outils. Une telle conception ne confère pas un rôle probatoire à la mesure quantitative. La fréquence n'est pas le résultat. Elle offre une prise à l'analyste, qui s'en empare (ou non) comme d'un marqueur possible. [47, LEJEUNE et BENEL] »

Autre garde-fou possible, le retour aux textes. Afin d'éviter le dérapage interprétatif, tout résultat nécessite d'être confronté à sa remise en contexte. Certes, les segments répétés, les concordanciers constituent déjà de bons moyens de contextualiser les termes extraits, de voir rapidement quel est leur contexte avoisinant. Mais « l'extraction de ces fenêtres graphiques ne respecte pas totalement les frontières naturelles des zones textuelles [97, HABERT, NAZARENKO et SALEM] » et ne saurait donc se substituer au retour aux matériaux textuels d'origine. L'importance statistique d'un fait doit donc être rapportée à son milieu d'origine, pour vérifier que l'interprétation qui en est donnée est pertinente. Car l'analyse lexicométrique peut très facilement donner prise à une « mise en histoire » des résultats, à

[53] Est bruyant, pour Calliope, un terme qui ne participe pas à l'élaboration d'un cluster de termes coocurrents, en raison de poids informationnel trop faible.

un enchaînement apparemment logique d'éléments discrétisés. C'est pourquoi « les résultats fournis par le logiciel sont des pistes qui réclament un retour à la globalité et à la linéarité des textes et qui doivent être croisés avec d'autres types de faits pour une analyse complète. [97, HABERT, NAZARENKO et SALEM] »
Enfin, toujours dans la perspective de rapporter les résultats à leurs conditions d'élaboration, soulignons la nécessité de soigneusement documenter l'intégralité du processus d'analyse textuelle : du mode de collecte des données à la sélection des documents (et donc du rejet d'autres documents) lors de la constitution du corpus, des divers prétraitements effectués sur les données (nettoyage, normalisation) aux paramètres choisis au sein des outils, du choix de variables à la sélection de certains graphiques parmi tout l'éventail possible, etc. Seule une telle documentation pourra permettre de comprendre, d'une part, le processus d'élaboration des résultats mais aussi d'en revoir certaines modalités, en fonction des questionnements progressifs. Ces renseignements s'avèrent également essentiels dans le cadre d'une exploitation secondaire des données. Outre la capitalisation (patrimoniale) qu'une telle démarche représente, elle peut donner lieu à une diversification des analyses[54].

3.4.2.4 Visualisation des résultats : des artefacts à accompagner d'un discours critique

Bien que fort utiles, en ce qu'ils permettent de représenter synthétiquement la diversité lexicale, les graphiques produits par les outils lexicométriques recèlent de nombreux pièges qu'il s'agit de comprendre et d'expliciter. De même que la phase de traitement des données opérait des réductions et des manipulations du matériau textuel, la représentation graphique des résultats est un artefact, une construction dont la lisibilité demande à être questionnée. B. Dantier insiste sur le fait qu'une « représentation graphique, pour devenir efficiente et surtout pour ne pas constituer un obstacle parasite au traitement de l'information, réclame maintes précautions qui doivent être toujours recommencées et perfectionnées [74, DANTIER] ». Car si "un dessin vaut mille mots", elle implique également mille pièges ! La capacité de synthèse de l'image (d'innombrables données et de multiples variables sont en effet rassemblées au sein d'une image, projetées dans les dimensions de l'espace) permet de voir, globalement, des relations et des structures imperceptibles à la lecture, fragmentée et linéaire, du matériau textuel. Mais cette lecture synthétique qu'offre l'image, bien que spontanée, est faussement neutre. Le risque principal serait de croire que le résultat visuel représenterait une vision complète d'un objet (dans notre cas, « Le » vocabulaire mémoriel employé au sein des déclarations d'associations) alors qu'il ne s'agit que d'un artefact visualisant certaines données, elles-mêmes sélectionnées, traitées et organisées. Ainsi, M. Grandjean rappelle que « sans discours critique qui en explicite les sources (et la méthode dans le cas de visualisations complexes), cette visualisation n'a pas toute sa légitimité narrative ou explicative. [78, GRANDJEAN] » La lecture d'un graphique appelle donc un discours critique qui en expose les sources, les méthodes d'élaboration et une sensibilisation aux possibles biais de lecture. Comment interpréter les axes d'une AFC par exemple ? Comment choisir l'algorithme de spatialisation dans la visualisation d'un graphe ? Quelles échelles sont utilisées ? Que représente la distance séparant deux clusters, la taille ou la couleur de leurs liens ? Que représentent les termes dans un cluster ? Etc. Ainsi, dans le cadre de notre étude, il s'agira, lors de l'étude des clusters produits par Calliope, de se rappeler que le lexique indexé a fait l'objet d'une validation préalable de l'expert, que certains termes synonymes ont été regroupés en raison de leur proximité sémantique, etc. Sans ce discours critique, l'image perd sa crédibilité.
Cette difficulté redouble lors d'une profusion de graphiques, comme c'est le cas notamment avec Alceste. Une multitude de restitutions graphiques sont ainsi offerts à l'utilisateur (tableaux et graphiques multiples, possibilité de naviguer, de zoomer et de se

[54] Voir notamment DUCHESNE S., « Développement de l'analyse secondaire et des méthodes d'analyse qualitative□: une chance à saisir? », M. BRUGIDOU, M. DARGENTAS, D. LE ROUX et A.C. SALOMON (dir.), *Analyse secondaire en recherche qualitative□: enjeux pour les sciences humaines et sociales*, Editions Lavoisier, 2007.

déplacer, etc.) mais au risque de le perdre : Sur quoi se concentrer ? Que sélectionner ? Quel élément est pertinent pour l'analyse ?
L'image n'est donc jamais le reflet d'une « réalité » mais bien une construction fondée sur des choix qu'il s'agit d'expliciter au moment de son interprétation. L'image n'étant que la face visible des processus et des choix intégrés tout au long de la démarche, elle ne saurait être figée, définitive. Evoluant au gré des interrogations et du cheminement interprétatif – dont elle rend compte –, elle ne peut être qu'une « porte d'entrée [78, GRANDJEAN] », guidée par les données, pouvant ouvrir des perspectives inédites.

3.4.2.5 Analyse de données textuelles : ne pas sous-estimer le facteur temps

Tout comme le fonctionnement de ces outils n'est pas neutre, leur usage ne se résume pas à « presser un bouton ». Et s'il est vrai que les outils d'analyse de données textuelles permettent d'accéder à des corpus beaucoup plus volumineux, il est tout aussi vrai que l'utilisation de ces outils se voit limiter par certains facteurs, dont l'un est, de façon paradoxale, le facteur temps.
Le temps de traitement des données a été fortement réduit grâce à la croissance exponentielle de la puissance de calcul des ordinateurs mais pour augmenter par ailleurs… Le paradoxe n'est qu'apparent, dans la mesure où il s'agit de deux moments différents. Dans un cas, le traitement des données au moyen des ordinateurs et des programmes dédiés caractérise le temps de calcul et de traitement statistique par la machine (au sens de « *computarer* ») de millions de caractères. Dans l'autre, il s'agit des étapes situées en amont (prétraitements) et en aval (analyse et interprétation des résultats) de la « computation informatique » proprement dite, qui échappent à la machine et requièrent une implication importante de l'humain – donc du temps. Et, de manière quasi géométrique, le temps consacré aux deux étapes encadrant le traitement informatique croît proportionnellement au nombre de données traitées. Ainsi, chaque étape peut se révéler chronophage : choix des données, structuration, normalisation et homogénéisation des données, prétraitement(s) du corpus, tests de paramétrage et sélection de résultats pertinents, analyse et interprétation des résultats, retour aux textes, nouveaux tests, etc.
Chaque phase recèle ainsi ses propres difficultés : Quelles sont les sources disponibles et accessibles ? En quoi ces sources conviennent-elles aux problématiques du projet ? Quel est le mode de collecte des données ? Quelle était la finalité initiale qui a gouverné la collecte des données (en cas de réutilisation de données dans un autre contexte) ? La constitution du corpus est-elle la plus pertinente possible au regard de la finalité ? Les données sont-elles structurées ? Sont-elles structurées de manière uniforme dans le temps ? Se présentent-elles sous un format nativement numérique ? Sinon, quelle est la qualité de l'océrisation ? Les résultats de cette océrisation nécessitent-ils un nettoyage ? Le développement de scripts est-il alors nécessaire ? Quel est le degré d'homogénéité des données ? Quels outils choisir ? Quelles opérations sur les données sont proposées ? Quels sont les formats admis par les outils utilisés ? Le changement de format affecte-t-il les données ? Le traitement par les outils implique-t-il une mise en forme particulière ? Quels sont les divers paramétrages proposés par les outils ? En quoi la variation de paramètres affecte-t-elle les résultats ? Quelles sont les formes de restitution des résultats ? Sont-elles aisément lisibles, interprétables ? Etc. Nous voyons donc que, loin de se résumer à être un presse-bouton, l'usage d'un outil d'analyse de données, qui plus est textuelles, traduit une véritable démarche, englobant aussi bien l'élaboration d'hypothèses et de problématiques, la vérification de l'adéquation des sources et données à ces problématiques que la disponibilité de ressources et de compétences, etc.
Chaque étape, dans la mesure où elle peut affecter le processus dans son intégralité, doit être clairement analysée, discutée, revue, modifiée, etc. Un projet peut ainsi voir son périmètre initial redéfinit en raison d'une mauvaise estimation, voire absence, de la qualité des données à traiter – qualité qu'il est, par ailleurs, toujours difficile à évaluer avant de s'y être confronté concrètement, ie avant d'avoir débuté le projet ! L'ensemble de ces opérations, et des choix sur lesquels elles reposent, implique donc nécessairement de disposer du temps requis.

3.4.2.6 De multiples compétences et connaissances à assimiler : Vers une division du travail ?

Dans un projet d'analyse de données textuelles, et de manière plus générale, dans tout projet d'humanités numériques, il ne faut pas non plus sous-estimer le coût d'apprentissage des compétences et connaissances. S'il ne s'agit pas réellement pour l'historien (ou le sociologue) de devenir programmeur[55], l'importance des connaissances et compétences à acquérir dans ce type de démarche est néanmoins considérable. Car l'usage des outils lexicométriques ne saurait se réduire à une simple prise en main mais requiert l'assimilation des principes statistiques et linguistiques sur lesquels ils sont fondés et qui, comme nous l'avons vu, influencent les résultats produits. Seule l'assimilation de leurs principes de fonctionnement, leur maîtrise technique et une compréhension des restitutions visuelles, permettra de ne pas en dénaturer l'usage, ie de ne pas les transformer en simples « presse-bouton ». Savoir ce qu'il est possible de faire (et de ne pas faire), ce qu'il est possible d'obtenir (et de ne pas obtenir) requiert un long temps d'apprentissage, composé d'essais et d'erreurs, de retours en arrière, de lecture, de discussions, de digestion, etc. Et ce recul a un coût. Ce constat fait d'ailleurs l'objet de l'une des conclusions du ThatCamp 2012[56].
Etant donné l'ampleur des savoirs et savoir-faire requis, l'option envisagée par certains consisterait à promouvoir des démarches collaboratives entre divers métiers. Ainsi, E. Ruiz note que « l'une des solutions envisageables consiste dans le développement des démarches collectives. C'est probablement l'une des transformations majeures qu'est appelée à connaître le métier d'historien grâce au numérique : la facilitation des démarches collaboratives de grande ampleur laisse entrevoir des possibilités inédites d'exploitation et d'analyse de corpus considérables. [21, RUIZ] » Ce dialogue entre métiers, en vue d'une réalisation commune, serait fondé sur une culture commune, des repères communs : pour le chercheur, il s'agirait moins de savoir coder que d'avoir une connaissance des enjeux liés à l'utilisation des techniques numériques et, pour l'ingénieur, d'être sensibilisé aux méthodes scientifiques et de connaître les problématiques visées.

En guise de conclusion à cette partie sur la sensibilisation aux biais possibles de l'analyse textuelle et les manières d'y remédier, nous présentons une synthèse des principes gouvernant l'analyse de données, élaborés par [54, OLLIVIER] :
- Absence de neutralité des outils : les logiciels sont de véritables « épistémologies embarquées » ;
- Ne jamais séparer les résultats obtenus des conditions de leur production ;
- Aucune solution clé en main automatisée n'existe ;
- Les logiciels sont au service des questionnements sociologiques ;
- L'usage d'un logiciel est inséparable d'une réflexivité, organisant les va-et-vient entre corpus, algorithme de traitement, hypothèses, questionnements et problématiques sociologiques.

Ce n'est qu'une fois ces principes connus et admis, que les outils d'analyse de données textuelles peuvent dévoiler tout leur potentiel et nourrir le raisonnement scientifique.

[55] En référence à la déclaration faite par E. Le Roy Ladurie, en 1967.
[56] Voir le texte sur http://tcp.hypotheses.org/336.

4 Panorama des outils d'analyse de données textuelles

Avant de présenter les logiciels utilisés dans le cadre de cette étude (Calliope et Alceste), nous allons dresser un rapide panorama des outils d'analyse textuelle[57]. Le nombre et les fonctions de ces outils représentent un vaste ensemble, dans lequel il est difficile de repérer. C'est pourquoi nous ne tenterons pas ici procéder à un inventaire détaillé de ces logiciels ni d'en préciser les présupposés épistémologiques, mais plutôt, en nous basant sur des classifications déjà élaborées, d'en présenter les points saillants, regroupés en grandes familles. Précisons néanmoins que cette présentation n'a qu'une finalité pratique : chaque outil a pu évoluer d'une spécificité d'origine vers d'autres domaines et intégrer des fonctionnalités qui, au départ, en étaient absentes ou bien, présenter des spécificités (analyse sémantique par exemple) qui n'excluent pas d'autres fonctionnalités (calculs statistiques). De multiples approches du matériau textuel peuvent ainsi caractériser ces outils, qui tendent à devenir de véritables couteaux suisses. D'où la difficulté à les départager.
Une première ligne de partage pourrait être établie entre dominante statistique ou linguistique des outils, une autre distinguerait analyse textuelle proprement dite (statistiques textuelle) et analyse de contenu. Une autre typologie, élaborée par C. Lejeune[58], repose quant à elle sur la caractérisation des outils au moyen de cinq fonctionnalités (lexicométrie, concordances, automates, outils réflexifs et dictionnaires), distribuées selon deux axes (montrer / calculer ; explorer / analyser). Notons simplement que l'ensemble de ces outils manipule, de manière diversifiée, les données textuelles.

4.1 Classification générale des approches et outils

Nous retiendrons ici l'inventaire critique de J. Jenny[59], établi en 1997 mais qui demeure une référence incontournable :
- Les **approches lexicométriques[60]**, issues de la statistique d'analyse factorielle et de classification automatique de type « benzécriste », qui consiste à comparer des profils lexicaux des corpus. Cette approche est prolongée par la « logométrie[61] », qui complète la statistique lexicale par des analyses grammaticales et syntaxiques, de façon à appréhender le discours et non plus simplement le lexique.
> *Spad-T*, *Lexico*, *Alceste*, *Hyperbase*, *Iramuteq*, etc.
- L'**analyse socio-sémantique de contenu thématique**, qui procède par segmentation du corpus en unités de signification pertinentes et par catégorisation multidimensionnelle conforme aux grilles d'analyse conceptuelle spécifiques des recherches, et par recours éventuel à des méthodes statistiques.

---

[57] Sont compris sous la désignation « analyse textuelle » les outils relevant de la statistique textuelle et ceux de l'analyse de contenu et de discours.
[58] Voir LEJEUNE C., « Montrer, calculer, explorer, analyser. Ce que l'informatique fait (faire) à l'analyse qualitative », *Recherches qualitatives*, 2010, n° 9, coll. « Les actes », p. 15-32.
[59] Pour une présentation détaillée de cet inventaire critique et des postulats épistémologiques sur lesquels les outils se fondent, voir JENNY J., « Méthodes et pratiques formalisées d'analyse de contenu et de discours dans la recherche sociologique française contemporaine. Etat des lieux et essai de classification », *Bulletin de méthodologie sociologique (BMS) [En ligne]*, 1997, n° 54, p. 64-112.
[60] FALLERY B. et F. RODHAIN, « Quatre approches pour l'analyse de données textuelles□: lexicale, linguistique, cognitive, thématique », Montréal, AIMS, 2007., désignent, quant à eux, ces approches d'« analyse lexicale » et M. DE SAINT-LÉGER de « textométriques ».
[61] Voir MAYAFFRE D., « De la lexicométrie à la logométrie », *Astrolabe*, 2005, p. 1-11. et MAYAFFRE D., *Vers une herméneutique matérielle numérique. Corpus textuels, Logométrie et Langage politique*, Université Nice Sophia Antipolis, 2010.

> *Modalisa-Interv*i*ews*, *Sphinx-Lexica*, etc.
- L'**analyse automatique des réseaux de mots associés**[62], issue des paradigmes des représentations sociales en psychologie sociale et de l'analyse des réseaux en sociologie, qui vise à « re-présenter », via des méthodes de classification automatique, des configurations cognitives liées à un ou plusieurs thèmes, considérées comme cachées sous la surface textuelle.
> *Réseau-Lu*, *Leximappe*, *Calliope*, *Evocation*, etc.
- Les **courants d'analyse propositionnelle et prédicative du discours**[63], au contact des paradigmes cognitivo-discursifs de la psycholinguistique, qui décrit les logiques de construction progressive de tout univers référentiel cohérent, avec la notion de « schéma causal », ainsi que les finalités ou intentions de chaque mise en scène langagière particulière, avec différents « opérateurs argumentatifs ».
> *Tropes*, etc.
- Les **logic**ie**ls d'ingénierie textuelle**[64], à dominante d'audit textuel ou de documentation-communication, incluant les logiciels généralistes d'analyse d'enquêtes sociologiques (les plus proches des CAQDAS[65] anglo-saxons), qui comportent des modules d'analyse textuelle notamment pour le traitement des questions ouvertes. Ces outils permettent d'organiser et de retrouver l'information à partir de marqueurs définis et apposés manuellement par le chercheur (annotation, codage du texte).
> *Atlas-Ti*, *NVivo*, *MAXQDA*, etc.
- Et enfin, les **logic**ie**ls dédiés à des problématiques de recherche particulières**, mais susceptibles d'applications hors de leur domaine initial.
> *Civilité*, *Coconet*, *Prospero*, etc.

Comment choisir un outil ?
Etant donné la diversité et la multitude d'outils disponibles, il n'est pas aisé de savoir lequel choisir. De fait, la sélection d'un logiciel s'effectue en fonction de plusieurs critères : la nature textuelle du corpus que l'on traite (données d'enquête, entretiens, questions ouvertes, textes littéraires, articles journalistiques, documents institutionnels, etc.), son volume et son type (taille du corpus, degré d'homogénéité, formats des données), la problématique étudiée et les types de résultats attendus (classer des textes ou des fragments de textes, extraire des informations, effectuer une synthèse, inventaire des thèmes, enrichir un corpus de commentaires, l'annoter, etc.[66]), les moyens mis à disposition, leur coût d'entrée (prix du logiciel, coût d'apprentissage), etc.

---

[62] Pour LEJEUNE C., cette catégorie d'outils rassemble les « automates » qui procèdent à des catégorisations automatiques, basées sur un calcul de cooccurrences ou une analyse factorielle. Des outils tels que Spad-T et Alceste, qui possèdent cette fonctionnalité, peuvent également être rangés dans cette catégorie Pour M. DE SAINT-LÉGER, ils opèrent une « classification automatique ».
[63] Pour M. DE SAINT-LÉGER, cette approche est qualifiée de « sémantique », dans la mesure où elle propose l'élaboration d'ontologies et de thésaurus et des techniques de classification sémantique.
[64] Soulignons, ainsi que le remarquent B. GARNIER et M.-F. GUÉRIN-PACE, *in* GARNIER B. et F. GUÉRIN-PACE, *Appliquer les méthodes de la statistique textuelle*, Paris, CEPED, coll.« Les collections du CEPED - Les clefs pour », 201O, que les logiciels d'aide à la lecture d'entretiens (NVivo) ou à la post-codification (Sphinx-Lexica) ne sont pas des logiciels de statistique textuelle.
[65] CAQDAS est l'acronyme de « *Computer Assisted / Aided Qualitative Data Analysis Software* ». Ces outils assistent le chercheur dans sa lecture d'un corpus textuel, en lui proposant tout un éventail de fonctionnalités, telles que des modules de codage et d'annotation, d'interrogation des textes ou encore de représentation de réseaux. Pour en savoir plus, voir notamment http://www.surrey.ac.uk/sociology/research/researchcentres/caqdas/. Pour FALLERY B. et F. RODHAIN, ces outils fournissent une analyse thématique assistée et constituent, pour LEJEUNE C., des outils réflexifs.
[66] Voir BRUGIDOU M., C. ESCOFFIER, H. FOLCH, S. LAHLOU, D. LE ROUX, P. MORIN-ANDRÉANI, et G. PIAT, « Les facteurs de choix et d'utilisation de logiciels d'Analyse de Données Textuelles », *op. cit.*

Nous présentons ci-dessous un tableau rassemblant quelques-uns des facteurs de choix d'un logiciel d'analyse textuelle[67].

	Analyse lexicale / Lexicométrie	**Analyse de discours / Analyse sémantique**	**Analyse de contenu (CAQDAS)**
Outils	Alceste, Calliope, Hyperbase, Lexico	Prospero, Tropes	Atlas-Ti, NVivo
Cadre méthodologique	Exploratoire	Exploratoire	Exploratoire
Axe temporel	Instantané (Alceste) Longitudinal (Calliope)	Instantané	Instantané Longitudinal
Objet d'analyse	Groupe	Individu	Projet
Taille du corpus[68]	Importante	Limitée	Importante
Homogénéité du corpus[69]	Faible	Forte	Faible
Moment de l'analyse statistique[70]	Ex-ante Contrôle ex-post	Ex-ante	Contrôle ex-post

4.2 Présentation des deux logiciels, Calliope et Alceste[71]

Les outils retenus ici, Calliope et Alceste, sont tous deux fondés sur des méthodes statistiques qui ont en commun les finalités suivantes :
- Déterminer statistiquement comment les éléments d'un corpus s'organisent, au moyen d'analyses multivariées (classification hiérarchique et AFC) ;
- Réduire l'arbitraire dans la description du corpus (limiter les perceptions subjectives, par une appréhension globale et non linéaire des textes) ;
- Mettre en évidence les informations essentielles contenues dans un corpus (calcul de spécificités lexicales et comparaison de profils lexicaux).

Malgré la proximité des deux outils quant à leur finalité générale, ils se distinguent par les méthodes employées et les résultats obtenus. Si la méthode implémentée au sein de ces outils repose en effet sur une technique de classification hiérarchique, celle-ci diffère selon le logiciel considéré.

[67] Ce tableau, simplifié et légèrement modifié, est issu de FALLERY B. et F. RODHAIN, « Quatre approches pour l'analyse de données textuelles□: lexicale, linguistique, cognitive, thématique », *op. cit.*
[68] L'analyse lexicale permet de traiter plusieurs milliers de documents (Alceste), l'analyse de discours une centaine.
[69] Ce critère tient compte de la diversité des sources, de la distinction entre types de textes, etc.
[70] Dans une approche statistique *ex-ante*, le traitement des données « guide » l'interprétation (ou plutôt lui sert de prétexte pour d'autres questionnements). L'approche *ex-post*, quant à elle, permet de contrôler la validité de catégories établies au préalable par le chercheur (les calculs statistiques n'interviennent qu'à la fin du processus, comme pour les CAQDAS).
[71] La présentation des outils donnée ici se veut générale, elle ne décrira donc pas leur méthode de façon détaillée.

4.2.1 Méthodes de classification : classifications hiérarchiques ascendante et descendante

Dans le domaine de l'analyse de données, notamment textuelles, la notion de classification[72] désigne les diverses méthodes de regroupement des données (*data clustering* en anglais), au moyen d'algorithmes de classification. Il s'agit de procéder au regroupement de données, via des méthodes statistiques d'analyse des données, en différents « paquets » homogènes (classes). Les données de chaque sous-ensemble partagent des caractéristiques communes, qui correspondent le plus souvent à des critères de proximité (similarité ou dissimilarité). En d'autres termes, le clustering cherche à élaborer des classes telles que :
- Les différences intra-classe soient minimales pour obtenir des clusters ;
- Les différences inter-classe soient maximales afin d'obtenir des sous-ensembles bien différenciés[73].

Les méthodes de classification hiérarchiques se décomposent en deux types, selon le point de départ de l'analyse : l'analyse débute soit au niveau élémentaire où chaque donnée est considérée individuellement (il s'agit de la « classification ascendante hiérarchique », utilisée par Calliope), soit par la prise en compte de l'ensemble des données (il s'agit alors de la « classification descendante hiérarchique », implémentée dans Alceste).

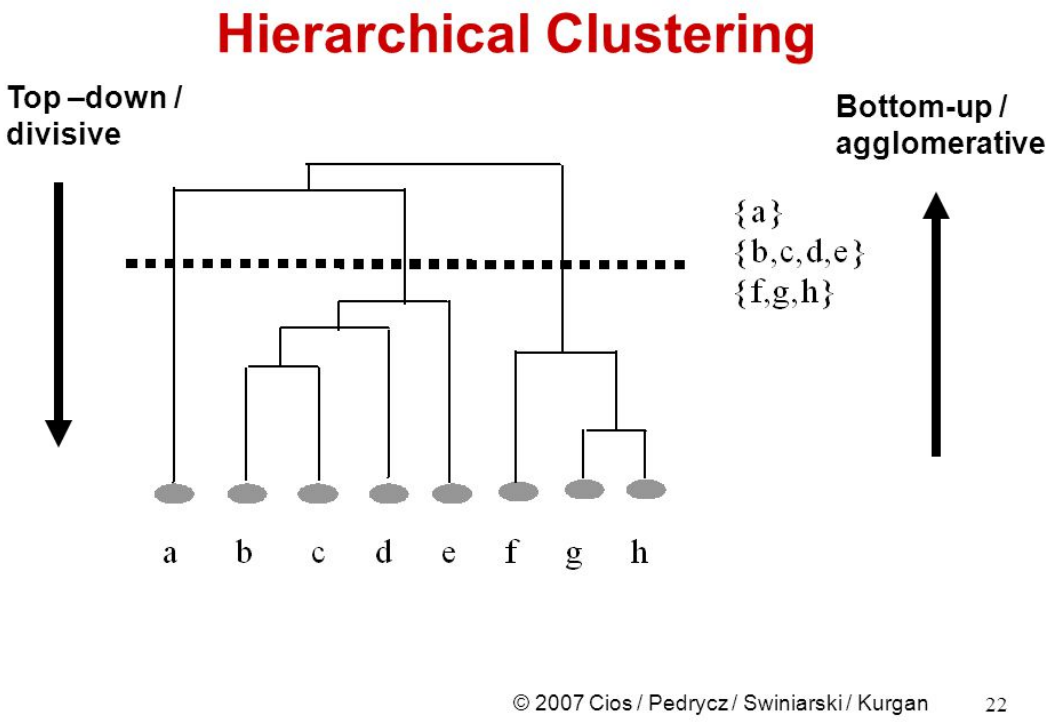


Figure 2 – Classification hiérarchique ascendante et descendante

[72] Nous n'abordons pas ici un troisième type de classification, la « classification par partitionnement » (classification non hiérarchique) dans laquelle, contrairement aux CDH et CAH, le nombre de classes est fixé à l'avance.
[73] Page Wikipédia, « Regroupement hiérarchique » : https://fr.wikipedia.org/wiki/Regroupement_hiérarchique

4.2.1.1 Classification ascendante hiérarchique (CAH) : agrégation des données par itération

Les algorithmes d'agrégation hiérarchique ascendante démarrent au niveau élémentaire, c'est-à-dire au niveau de chaque donnée (individu). Après une mesure des distances entre individus, elle procède alors par étapes successives, chacune d'elles consistant à regrouper les deux objets les plus proches. A la fin de chaque étape, les distances entre le groupe nouvellement créé et le reste des objets sont recalculées. Ce processus est réitéré jusqu'à ce que tous les objets aient été réunis dans un seul groupe.
A noter que cette méthode de classification est très utilisée en analyse de données, et notamment par Calliope. L'élaboration des clusters peut reposer sur plusieurs mesures de distance[74] entre éléments. Celle implémentée au sein de Calliope correspond à la méthode du « single link clustering » (ou agrégation par "le saut minimum" ou "du lien simple"), selon laquelle l'agrégation entre deux groupes se base sur la plus petite des distances entre deux groupes (plus petite distance inter-groupes).

4.2.1.2 Classification descendante hiérarchique (CDH) : découpage du corpus par dissemblance

Les algorithmes élaborant une classification hiérarchique descendante partent de l'ensemble entier de tous les individus, considérés comme une seule classe, puis procèdent par dichotomies successives, où les groupes constitués sont le plus dissemblable possible entre eux. A chaque subdivision, la distance entre éléments est recalculée, jusqu'à ce que tous les sous-ensembles obtenus soient réduits à un objet unique.
Selon M. Roux, « ce type d'algorithmes a eu peu de succès jusqu'à présent à cause des inconvénients majeurs qu'il présente. En effet, pour obtenir de bons résultats, il faudrait examiner à chaque étape toutes les dichotomies possibles pour n'en retenir qu'une, celle qui optimise un critère fixé à l'avance ». Malgré sa moindre popularité, c'est la méthode choisie par M. Reinert, le concepteur d'Alceste.

4.3 Présentation du logiciel Calliope

Le logiciel Calliope[75], conçu par Mathilde de Saint-Léger, est issu de préoccupations documentaires d'indexation de la littérature scientifique et technique (IST). Les premiers logiciels[76] développés dans ce domaine entendaient fournir une aide à l'indexation et à l'interrogation de bases de données bibliographiques : la terminologie de Calliope en porte encore la trace[77]. La finalité de Calliope est de présenter un découpage thématique d'un ensemble documentaire, qui informe sur la visibilité des thèmes, leur maturité et leur degré

---

[74] D'autres méthodes de mesure de distance existent : comme la « hiérarchie du diamètre » ou du « lien complet » (« *complete link* »), basée sur la plus grande distance interne au groupe résultant ou encore, la « hiérarchie de la distance moyenne » (« *average link* ») où la nouvelle distance vaudra la moyenne des distances antérieures. Pour plus d'information, voir ROUX M., *Algorithmes de classification [En ligne]*, Paris, Masson, 1985.
[75] Cette présentation provient essentiellement de la thèse de doctorat (1997, CNAM) de M. DE SAINT-LÉGER, « Modélisation des flux d'information scientifique et technique par le bruit : vers un suivi des domaines de la connaissance », 1997. En cas de référence à d'autres sources, nous les citerons.
[76] Calliope est l'héritier des logiciels Lexinet et Leximappe, développés par des sociologues (travaux de M. Callon, B. Latour, J.-P. Courtial notamment) du CSI (Centre de sociologie des innovations, CNRS / Ecole nationale supérieure des Mines) et des chercheurs et ingénieurs du CDST (Centre de documentation scientifique et technique du CNRS).
[77] « L'objectif des systèmes de gestion de stocks documentaires est de collecter, stocker puis restituer des documents issus de sources diverses. Pour atteindre cet objectif, il est d'usage de représenter un document dans une base documentaire par son titre, son résumé ou encore une liste de mots-clés (document secondaire) », *in* M. DE SAINT-LÉGER, *op. cit*. Les termes servant à l'indexation des documents sont appelés « descripteurs » (terminologie en usage au sein de Calliope).

de structuration. Une représentation schématique est ainsi élaborée où les documents homogènes sont agrégés et représentés par leurs principales thématiques (cooccurrence de termes). Une telle classification doit permettre de résumer le corpus, d'un point de vue thématique.
En raison de ses origines, Calliope combine approches documentaire et sociologique, au moyen de la méthode dite des « mots associés ». Cette méthode permet de représenter le corpus au moyen de classes de documents dont le vocabulaire cooccurre, ie dont les mots sont associés.

4.3.1 Méthode des mots associés

Cette méthode entend répondre, comme le rapporte M. de Saint-Léger dans sa thèse, à une question posée par W. A. Turner : « Comment faire apparaître dans les différents documents qui circulent (articles scientifiques, brevets, rapports, modes d'emploi...) les mises en relation opérées par les acteurs ? La réponse à cette question compliquée est d'une simplicité biblique : en identifiant les mots associés à l'intérieur de chacun de ces documents et en comptabilisant leurs associations. »
« La méthode des mots associés[78] est ainsi fondée sur la mise en évidence de réseaux d'associations entre les descripteurs d'un corpus documentaire. Chaque réseau (ou agrégat ou encore cluster) ainsi constitué est une unité informationnelle, une thématique. Leximappe [ici, Calliope], qui est un logiciel fondé sur cette méthode, positionne ces dernières les unes par rapport aux autres en fonction de leur interconnectivité (ou liens externes) et de leur cohérence interne (ou liens internes) sur un diagramme dit stratégique. En comparant deux cartes représentant deux flux d'information successifs d'un même domaine de recherche, on prend connaissance de la variation du contenu des agrégats ainsi que de leur migration sur les diagrammes stratégiques successifs. »
L'originalité de cette méthode consiste, selon J. Jenny, à « définir les "acteurs/actants" précisément par leur "profil d'association", c'est-à-dire la liste des mots auxquels ils sont associés (Latour) et à définir le contenu textuel comme "le réseau des associations opérées par le texte entre les acteurs qu'il met en scène" (Teil). Système dynamique (chaque acteur créant ou détruisant des relations avec d'autres acteurs, les réseaux se transforment sans cesse) et système ouvert (à tout nouvel acteur, controverse, consensus ou conflit). [36, JENNY] »
L'analyse des mots associés se propose ainsi de transformer une information quantitative (les associations entre descripteurs au sein d'un corpus de textes) en information qualitative (la structure du réseau des associations en émergence).

Cette méthode permet ainsi, non seulement de repérer, au sein d'un corpus textuel, des classes thématiques, qui sont construites en fonction de la proximité de leurs termes (calcul de cooccurrences), mais surtout d'en analyser la dynamique[79]. Car s'il est possible d'établir la représentation thématique d'un corpus à un moment donné, il est également possible d'en

[78] Selon J.-P. Courtial, « la méthode d'analyse des mots associés définit l'association entre deux mots-clés comme le produit des probabilités d'avoir un mot clé quand on a l'autre ». Le logiciel évalue la force des liens entre paires de termes et procède à leur regroupement au sein de clusters. Lorsqu'un cluster est constitué, la méthode poursuit sa construction des autres clusters. « Les autres paires contenant les mots retenus dans un cluster ne sont plus prises en compte pour la construction des autres clusters. Les liens qu'on ne peut plus prendre en compte sont cependant conservés pour le calcul des liens externes de thème à thème. », *in* COURTIAL J.-P. et L. KERNEUR, « La méthode des mots associés, outil d'analyse du changement social », *Histoire & Mesure*, 1997, vol. 12, n° 3/4, p. 251‑270.
[79] Ainsi, « dans l'esprit du processus inventif ou innovant, l'analyse des mots associés décèle des liens faibles, à l'origine de bouleversements, là où une analyse factorielle ne met en évidence que les aspects les plus stabilisés – donc les moins changeants – d'une structure de corrélations. Toutes les méthodes d'analyse des structures stables au sein d'espaces d'objets supposés totalement liés entre eux, comme les analyses factorielles ou les classifications en espace euclidien, sont, de ce point de vue, impuissantes à rendre compte du changement social, sauf à l'insérer dans des tendances lourdes. », *in Ibid.*

mesurer l'évolution temporelle, par une comparaison entre les cartes thématiques produites à divers instants.

4.3.2 Mode opératoire de Calliope

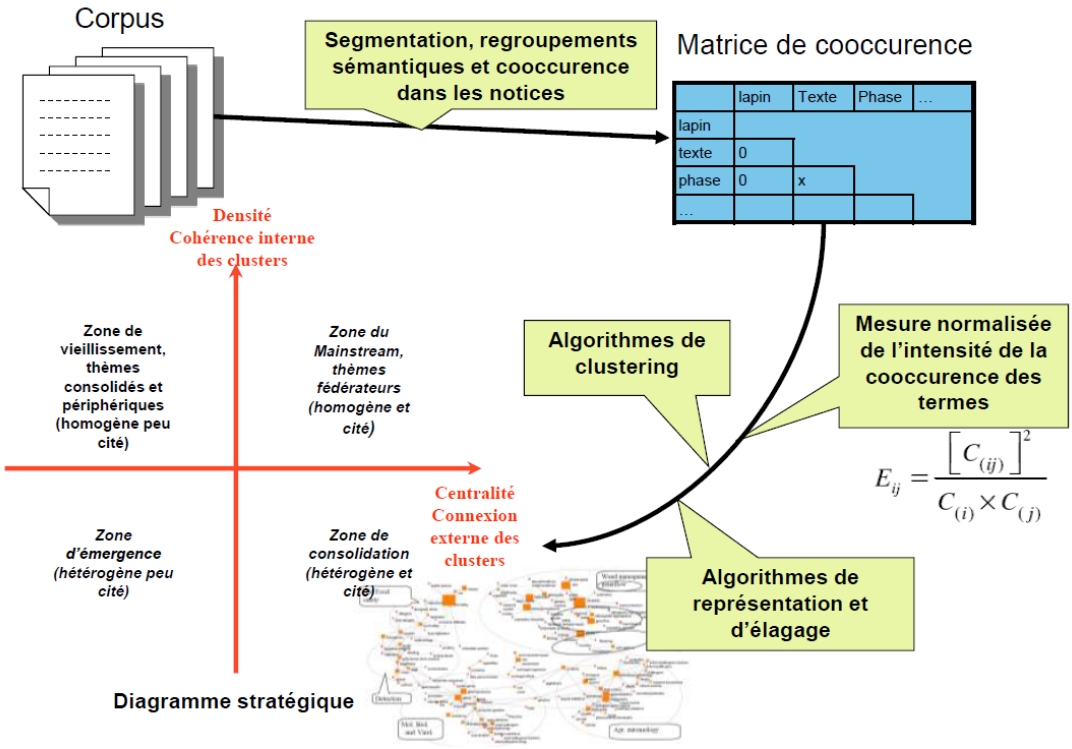


Figure 3 – Mode opératoire de Calliope[80]

Après l'extraction et la validation[81] du lexique, Calliope procède au regroupement des termes du corpus textuel (constitution de clusters), au moyen d'une classification ascendante hiérarchique.

Les clusters de termes sont positionnés sur un diagramme stratégique et représentés par leur "mot central" suivant deux variables, la centralité et la densité[82] :
- La **centralité** reflète le pouvoir de liaison d'un cluster, qui est d'autant plus fédérateur dans un domaine qu'il est central, c'est-à-dire qu'il est relié à d'autres pôles thématiques. Elle définit au niveau de chaque cluster la notion d'interconnectivité entre classes.
- La **densité** reflète la cohérence et la structuration interne d'un cluster, qui est d'autant plus dense que l'association entre ses termes est forte.

---

[80] Schéma issu de G. Ollivier, « Panorama critique des analyses textuelles informatisées en SHS », *op. cit.*
[81] La validation du lexique fait l'objet d'une présentation plus approfondie au sein du chapitre « Méthodologie » de ce mémoire. Disons simplement que la méthode employée par Calliope est semi-automatique en ce sens qu'elle implique une validation du lexique d'annotation du corpus par l'expert. En cela, elle se distingue d'Alceste.
[82] « Chaque composante connexe ou thème peut ainsi être caractérisé par : a) sa centralité, c'est-à-dire la somme des liens des mots qui la composent avec les autres mots (liens externes) ; b) sa densité, c'est-à-dire la valeur moyenne des liens entre mots du thème (liens internes). », *in* COURTIAL J.-P. et L. KERNEUR, « La méthode des mots associés, outil d'analyse du changement social », *op. cit.*

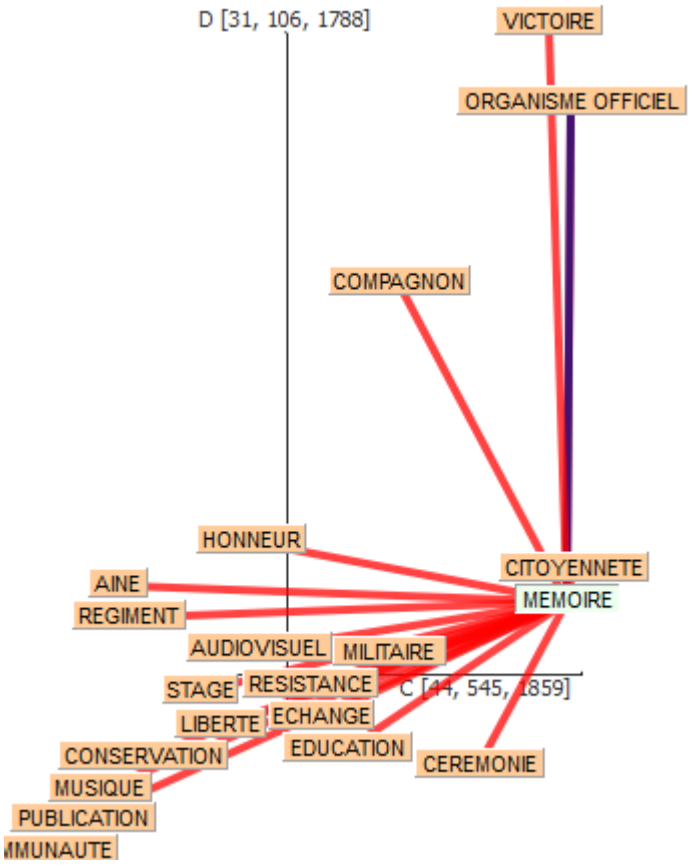


Figure 4 – Exemple de Diagramme stratégique (2010) : liens externes (centralité)

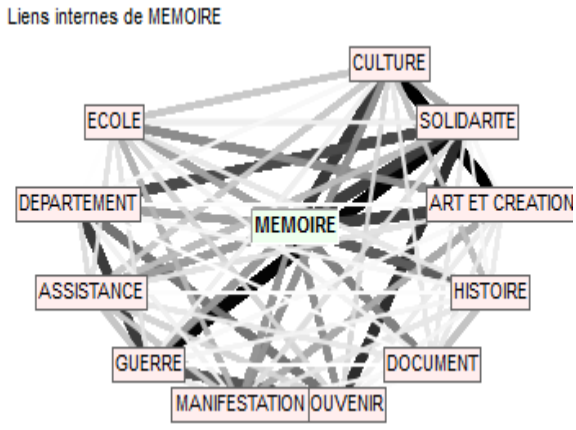


Figure 5 – Exemple de cluster (2010) : liens internes du cluster « Mémoire » (densité)

La lecture du diagramme stratégique et des clusters donne des indications sur la structuration et le positionnement des thématiques constitutives d'un corpus. Ainsi, le nombre de liens internes (et eur épaisseur et leur couleur[83]) d'un cluster renseigne sur sa

[83] Une gradation visuelle de l'importance des liens est exprimée au sein de Calliope : de manière décroissante, les liens vont du noir au blanc, en passant par différentes nuances de gris, et d'épais à très fins.

cohésion thématique : plus le nombre est élevé, plus les liens de cooccurrences sont forts, et plus le cluster est thématiquement homogène.

A titre d'exemple, comparons le cluster « Mémoire » (ci-dessus, en 2010) à celui de 1984 (ci-dessous) : le faible nombre de liens (associé à des traits à dominante blanche et d'épaisseur fine) est l'expression d'une thématique très peu structurée, dont les termes bien que cooccurrents, partagent peu de liens entre eux. A l'inverse, le cluster « Mémoire » de 2010 est beaucoup plus dense (liens plus nombreux et plus épais) : ce qui correspond à un plus grande homogénéité thématique.

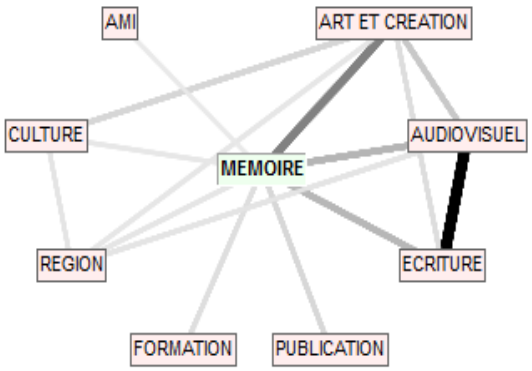


Figure 6 – Liens internes du cluster « Mémoire » (1984)

4.3.3 Diagramme stratégique, une vue synthétique des thématique d'un corpus

De même, le diagramme stratégique apporte un certain nombre d'informations sur les rapports qu'entretiennent les thèmes du corpus, en fonction de leur positionnement au sein des quadrants :

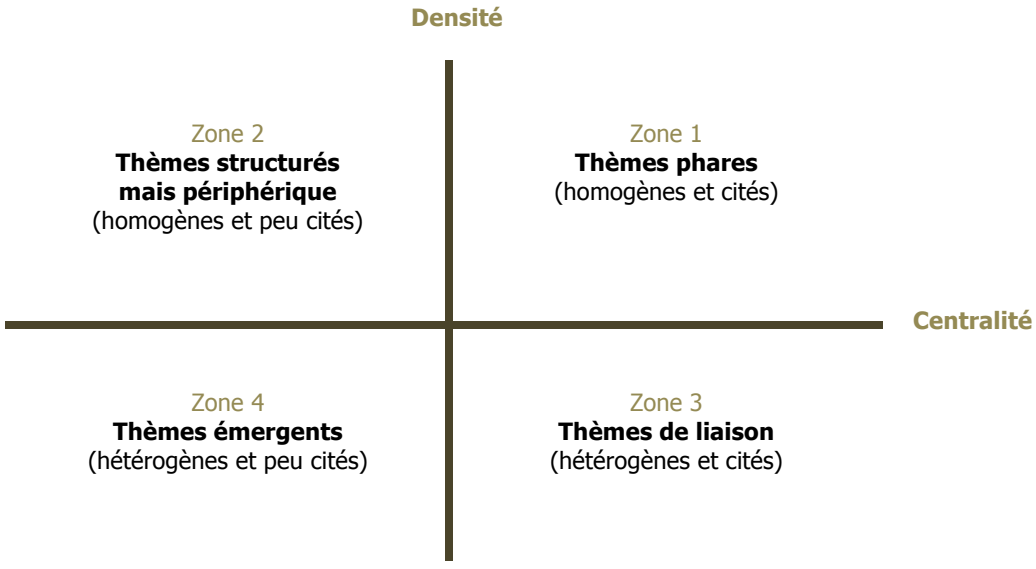


Figure 7 – Quadrants du diagramme stratégique

Ainsi, le diagramme fournit une représentation visuelle des divers types de thèmes du corpus :
- **Thèmes phares ou stratégiques** du corpus, au fort pouvoir structurant et fédérateur (quadrant 1) : les clusters relevant de ce quadrant sont à la fois fortement structurés (très denses) et entretiennent de nombreux liens aux autres clusters du corpus.

- **Thèmes structurés mais périphériques** (quadrant 2) : les clusters de ce quadrant sont relativement denses mais ont peu de liens aux autres thématiques du corpus (d'où la désignation de ce quadrant par l'expression « tour d'ivoire »).
- **Thèmes de liaison** (quadrant 3) : les liens entre les termes des clusters de ce quadrant sont lâches mais nombreux aux autres clusters du corpus. Ces thèmes, fréquents, apparaissent davantage en lien à d'autres thèmes qu'ils ne forment un ensemble homogène en eux-mêmes.
- **Thèmes émergents** (quadrant 4) : les thématiques de ce quadrant sont à la fois hétérogènes et peu citées au sein du corpus. Cette zone signale des thèmes en émergence, dont le poids informationnel pourrait se renforcer dans le temps (accroissement des liens internes et externes).

4.3.4 Visualisation des dynamiques lexicales

Calliope propose une analyse des dynamiques lexicales et thématiques d'un corpus, sur une période temporelle donnée ou entre documents publiés à la même période. L'analyse des tendances mesure ainsi les variations statistiques du poids informationnel des mots d'un sous-corpus à l'autre (même source, même type de documents, etc.).
Une comparaison peut également être établie entre deux corpus différents (différence de sources, types de documents, degré de complexité langagière variable, etc.).
Il est donc possible de comparer le contenu de plusieurs corpus entre eux :

- Comparaison du contenu de plusieurs corpus, indépendamment de toute notion temporelle : mesure des différences et similarités lexicales entre corpus (groupes de termes communs ou divergents).
- Par exemple, comparaison des stratégies communicationnelles à l'œuvre dans les discours politiques et publicitaires.
- Comparaison temporelle du contenu de deux ou plusieurs sous-corpus traitant d'un même thème et issus des mêmes sources : mesure de l'évolution temporelle d'un lexique.
  Par exemple, analyse temporelle des thématiques mémorielles au sein des déclarations d'associations au JO (notre étude).

4.3.5 Analyse chronologique

L'analyse chronologique que permet Calliope compare les positionnements des clusters au sein de diagrammes stratégiques élaborés à diverses périodes. Le principe repose sur la structure invariable des diagrammes stratégiques. En effet, contrairement aux clusters, qui fluctuent dans le temps en fonction de la proximité de leurs termes, la composition du diagramme stratégique présente une identité permanente, pérenne, en ce que ses zones désignent toujours, quelle que soit la période considérée, les mêmes types de thèmes (stratégiques, émergents, de liaison ou périphériques). Il est alors possible, en comparant la position qu'occupent des thématiques dans un diagramme élaboré à un temps t à celles qu'elles occupent dans un diagramme à un temps t+1 d'en déduire leur évolution temporelle.
Cette représentation passe de l'analyse de la participation des termes aux clusters à celle de leur participation au diagramme stratégique. Quel que soit le moment étudié, la zone 1 rassemblera toujours les thèmes phares du corpus, la zone 2, les thèmes de la tour d'ivoire, etc. Ainsi, un thème qui passe, à une date donnée, de la zone 4 (thèmes émergents) à la zone 1, à une date ultérieure, aura vu son pouvoir d'attraction[84] et de connectivité se renforcer dans le temps (écart temporel entre les deux dates considérées). Il est donc possible de suivre les évolutions thématiques d'un corpus textuel issu des mêmes sources,

[84] Le pouvoir d'attraction est la capacité d'un terme à constituer un réseau lexical thématique homogène.

de voir les glissements lexicaux qui se sont produits, les thèmes qui se sont substitués à d'autres.
Cette représentation est calculée selon le degré de participation des termes à une zone du diagramme (le pouvoir d'attraction d'un terme dans un quadrant doit être supérieur à celui qu'il peut avoir dans d'autres quadrants) : sont prises en compte la fréquence relative d'un terme (qui évalue son rang au sein d'un quadrant et limite l'« effet de taille » de la fréquence absolue), sa typicalité (un descripteur est d'autant plus typique d'un quadrant qu'il y est concentré, et donc absent des autres quadrants) et sa participation effective (qui est le rapport de la fréquence du terme à sa fréquence totale – ce qui exclut les termes qui ne seraient présents au sein d'un quadrant qu'en raison de leur apparition dans des textes dont les autres termes participent, eux, aux clusters du quadrant).

4.3.6 Variation temporelle du pouvoir d'attraction des termes

Ces mesures permettent, non seulement, de représenter visuellement les thèmes d'un corpus via la comparaison de diagrammes stratégiques mais offrent aussi la possibilité d'autres représentations graphiques (radars, courbes, histogrammes[85]) ou de listes de termes. Il devient possible de repérer des termes qui, insignifiants (d'un point de vue statistique) en début de période, voient leur poids informationnel croître sur la période, au point de dépasser un seuil de « remarquabilité »[86].

Le poids informationnel des termes pouvant varier au fil du temps, ils peuvent changer de catégories :
- **Terme stable** : le pouvoir d'attraction d'un terme dans la construction des réseaux d'associations est invariable dans le temps.
- **Terme émergent** : le pouvoir d'attraction d'un terme s'accroît dans le temps. Cette augmentation peut être l'indice de nouvelles thématiques, se structurant dans le temps et dont le vocabulaire est de plus en plus employé au sein des documents du corpus.
- **Terme déclinant** : le pouvoir d'attraction d'un terme décroît dans le temps et annonce alors des thématiques qui soit périclitent, soit s'atomisent en plusieurs thématiques (ce peut être le cas d'un terme générique relevant d'un domaine qui, en fonction du degré d'avancement des connaissances, par exemple, peut être remplacé par une pluralité de termes plus spécifiques : cette thématique, au sens large, persiste dans le temps mais elle n'est plus référée de la même manière, le vocabulaire la désignant a évolué en se précisant).
- **Terme fluctuant** : le pouvoir d'attraction d'un terme est instable sur une période (il passe au-dessous et en-dessous du seuil de « remarquabilité » sur la période).

4.4 Présentation du logiciel Alceste

Le logiciel Alceste[87], conçu par Max Reinert, est un logiciel d'analyse de données textuelles, fondé sur les méthodes de la statistique textuelle issue notamment de Jean-Paul Benzécri, qui cherche à rendre compte de l'organisation interne d'un discours, par l'extraction des structures signifiantes les plus spécifiques d'un corpus.
« La méthodologie Alceste entre dans le cadre général des recherches en analyse de données linguistiques (Benzécri, 1981 ; Lebart, Salem, 1989) et consiste principalement en l'étude des

[85] Ces moyens de visualisation de la trajectoire des termes remarquables permettent d'apprécier l'évolution individuelle de chaque terme et de la comparer à celle d'autres termes. La représentation d'une gradation des fluctuations est donc possible : ainsi, certains termes seront davantage émergents que d'autres.
[86] Ce seuil, variable sur une période, est fonction des caractéristiques statistiques propres à chaque sous-corpus.
[87] Le terme Alceste est l'acronyme de « Analyse lexicale par Contexte d'un Ensemble de Segments de Texte » (il était auparavant celui de « Analyse des Lexèmes Co-occurrents dans les Enoncés Simples d'un Texte »).

lois de distribution du vocabulaire dans un corpus. De ce point de vue, cette approche peut être considérée comme dérivée de l'approche distributionnelle de Z. S. Harris, au moins dans son objectif général puisqu'il s'agit *non pas de chercher le sens d'un texte mais de déterminer comment sont organisées les éléments qui le constituent*. [63, REINERT] » Alceste utilise donc les méthodes de la statistique distributionnelle comme un moyen de cerner la structure discursive d'un corpus et d'en établir la typologie via l'identification de classes de discours.
Pour ce faire, la méthode implémentée au sein du logiciel Alceste – et qui constitue sa principale originalité par rapport à d'autres outils –, s'appuie sur une classification descendante hiérarchique qui analyse la caractérisation des énoncés (via la distribution de leurs mots et cooccurrents) comme unité de base des univers de discours dont le corpus est constitué. Alceste permet, non pas tant de saisir la répartition des mots entre eux au sein d'un corpus que les associations particulières reliant mots et types d'énoncés : Alceste tente ainsi de capturer les différents univers de discours qui composent un corpus. Contrairement à de nombreux logiciels en analyse de données (dont Calliope), qui partent des unités élémentaires (termes) et de leurs cooccurrents pour former, par agrégation itérative (CAH), des clusters représentatifs, Alceste découpe le corpus global en unités de contexte (par la méthode de la CDH) pour en extraire et classer les énoncés typiques.

4.4.1 Mondes lexicaux

Cette méthode s'appuie sur la théorie des « mondes lexicaux », de M. Reinert, selon laquelle les textes sont le reflet de diverses postures énonciatives, que l'analyse statistique de la distribution du vocabulaire permet de saisir. En ce sens, Alceste ne cherche pas tant à révéler les thèmes d'un corpus ni à décrire l'objet qu'il représenterait qu'à « analyser les traces de l'activité discursive [63, REINERT] » au sein des textes. Cette activité se donne à voir au travers des énoncés qui, selon Reinert, portent davantage l'action d'énonciation elle-même que sa représentation. Le langage en acte serait ainsi un faire avant d'être un savoir sur ce faire. La signification d'un mot, pour être saisie, doit être rapportée à son usage et donc, à ses divers contextes d'utilisation : les mots, dans cette conception, issue de Wittgenstein notamment, ne désignent pas d'abord quelque chose de la réalité mais se fondent sur des règles contextuellement et socialement signifiantes, apprises et reproduites par mimésis (« jeux de langage »[88]).
Afin de circonscrire les limites inhérentes à l'atomisation d'un texte en ses unités élémentaires (le plus souvent nominales) qui, en détachant les mots de leurs contextes proches ôtent toute dimension discursive à l'analyse, Alceste se focalise sur les apparitions conjointes de termes au sein de « fenêtres » d'apparition, les unités de contexte.

4.4.2 Unités de contexte

Qu'est-ce qu'une unité de contexte, dans la terminologie d'Alceste ? Sommairement énoncé, une unité de contexte élémentaire[89] (UCE) est un segment de texte mais qui ne correspond pas

---

[88] Voir notamment L. WITTGENSTEIN, *Le cahier bleu et le cahier brun*, 1988 (1re édition, 1951), Gallimard ou *Investigations philosophiques*, 1988 (1re édition 1961), Gallimard, dans lesquels Wittgenstein montre que le discours ne peut être séparé de l'acte qui le porte ni du contexte dans lequel il s'insère, le langage n'est pas appris par monstration mais en en usant (ou plutôt la fonction ostensive du langage n'en constitue qu'un aspect, qu'un jeu de langage, mineur qui plus est). Cette fonction référentielle du langage découlerait d'une prééminence accordée aux noms sur les autres éléments langagiers. Or, Wittgenstein se demande quelle pourrait être la fonction référentielle du mot « et » ou de « mais » A noter que, si Alceste écarte les mots-outils lors de l'élaboration des classes de discours (leur fréquence trop importante parasiterait, de fait, toutes les classes), il est néanmoins possible d'étudier leur place dans les différents types de discours, par l'analyse de leur distribution au sein des classes. Voir notamment REINERT M., « Postures énonciatives et mondes lexicaux stabilisés en analyse statistique de discours », *Langage et société*, 2007, n° 121-122, p. 189-202.

[89] L'unité de contexte élémentaire constitue l'unité de base des calculs effectués par Alceste, au sein de laquelle le logiciel va dénombrer les mots réduits (lemmes) distincts. Elle se distingue de l'unité de contexte initiale (UCI) qui correspond aux segments de texte d'origine du corpus (dans notre cas, chaque déclaration d'association est une UCI). Ce peut être également les réponses de personnes

systématiquement – ni même souvent – à la notion de phrase ni à celle du paragraphe[90]. Une unité de contexte est constituée d'un ensemble de mots voisins dont la délimitation varie, en fonction notamment de la taille et du type de textes, de façon à ce que les résultats obtenus soient stables et représentatifs des discours (les classes de discours), ie indépendants de ces variations[91]. La taille déterminée pour une unité de contexte varie donc selon le type de phrases, de discours à l'œuvre au sein du corpus. Ainsi, la délimitation de la taille des unités textuelles de *La recherche du temps perdu* ou celle de nos déclarations d'associations ne saurait être identique – et proposer un seuil identique aurait peu de valeur représentative. Puisqu'il s'agit de saisir quantitativement des spécificités textuelles, discursives, la dimension de la largeur de la fenêtre[92] sur laquelle repose l'analyse n'est donc pas indifférente. Cette méthode propose, de cette façon, une manière de tenir compte des variations et des particularités discursives qui caractérisent différents types de textes à analyser.
Les UCE constituent la base de l'analyse effectuée par Alceste (détermination des différentes classes spécifiques aux discours). Après avoir regroupés ces unités de contexte élémentaire en unité de contexte (UC[93]), Alceste a recours à une méthode de classification[94] originale, la classification descendante hiérarchique, qui regroupe les énoncés du corpus en classes, en fonction de leur distribution lexicale interne (ie, le nombre de mots lemmatisés distincts que contient cette UC).

4.4.3 Classification descendante hiérarchique : élaboration contrastive des classes

Comme vu précédemment, ce type de classification considère, en début d'analyse, l'ensemble des données textuelles du corpus comme une seule classe puis procède à son découpage en deux, de façon itérative, en comparant le lexique de chaque unité textuelle. Cette méthode s'accomplit par dichotomies successives où les groupes constitués, à chaque étape, sont le

---

interviewées lors d'entretiens, des articles de journaux ou encore, les chapitres de livres. Une UCI est délimitée par l'introduction d'une ligne étoilée (identifiant et ensemble de variables illustratives qui permettront, ensuite, de faire des tris croisés) et se poursuit jusqu'à la prochaine ligne étoilée. De cette manière, une UCE ne peut chevaucher deux UCI.

[90] Ces notions, qui seraient identiques d'un type de textes à l'autre, ne permettent pas d'appréhender toute la diversité des discours possibles. C'est pourquoi la délimitation des UCE est un compromis entre forme syntaxique (ponctuation) et contraintes statistiques (dont le seuil peut par ailleurs être modifié par l'analyste), fonction du type de corpus traité.

[91] Pour assurer la stabilité des résultats, Alceste procède à deux classifications successives, en faisant varier le nombre de mots analysés distincts au sein de la fenêtre.

[92] A noter que la « fenêtre » prise comme unité de comparaison diffère entre Alceste et Calliope. Pour Alceste, comme nous venons de le voir, il s'agit de l'unité de contexte alors que pour Calliope, il s'agit du document. Cette différence peut avoir une grande influence sur les résultats : la distance mesurée entre termes cooccurrents peut alors grandement varier selon la taille du document initial. En effet, Calliope évalue cette proximité à l'échelle du document tandis qu'Alceste la mesure à l'échelle d'un segment de texte. Précisons néanmoins que cette différence joue peu dans le cadre de notre étude dans la mesure où, même si la taille des déclarations varie dans le corpus, elle varie peu. Ainsi, près de 70% des déclarations ont un contenu textuel composé de 100 à 400 caractères (espaces compris), et 94% entre 100 et 1 000 caractères. Notre corpus est donc relativement homogène du point de vue de la taille.

[93] Les unités de contexte consistent en l'union de plusieurs UCE, en fonction du vocabulaire partagé. A noter que l'UCE constitue la plus petite unité de contexte défini par Alceste, l'UCI, la plus grande et l'UC est comprise entre ces deux types d'unités de contexte.

[94] La classification hiérarchique (ou *clustering*) est une méthode de détermination de classes à partir de données, fondée sur des mesures de proximité ou de distance entre données. La classification hiérarchique ne présuppose pas un nombre de classes fixé à l'avance. En cela, elle s'oppose à la classification par partitionnement. La finalité de la classification est d'obtenir des classes le plus homogène possible.

plus dissemblable possible entre eux et où chaque groupe est le plus homogène en lui-même, puis affine ce découpage par une autre scission en deux, de façon réitérée[95].
Cette classification statistique regroupe les « phrases » du corpus étudié en fonction de la distribution du vocabulaire à l'intérieur de ces « phrases ». Le logiciel repère ensuite les formes réduites dans les différentes unités de contexte et les met en relation : le logiciel, par le calcul des liens de proximité / distance entre unités de contextes, peut ainsi relier les contextes qui ont des mots communs. Il croise les unités de contexte et la présence / absence de ces formes dans les UC. En d'autres termes, il forme des classes à partir des UC contenant les mêmes mots [30, DELAVIGNE]. Des énoncés sont donc considérés comme typiques de la même classe s'ils partagent un certain nombre de mots et les classes s'opposent en raison de la distinction de leur vocabulaire, de leurs énoncés. Alceste présente ensuite une liste de termes par classe, sélectionnés comme pouvant décrire une classe : ces termes sont ceux qui apparaissent le plus souvent au sein des énoncés typiques. Ils sont rangés selon un ordre de spécificité décroissant[96]. Il faut cependant souligner que ces termes peuvent apparaître dans d'autres énoncés que ceux de la classe qu'ils décrivent, voire même être plus fréquents dans une classe mais plus significatifs d'une autre. En effet, si un terme apparaît moins dans une classe mais cooccurre toujours avec les mêmes termes, son degré de typicité est alors plus important que celui qu'il peut avoir dans une autre classe, dans laquelle il est plus fréquent mais associé à des termes, des contextes plus variés. Nous verrons cela notamment lors de l'étude des mots-clés mémoriels, lors de la présentation des résultats obtenus avec Alceste.

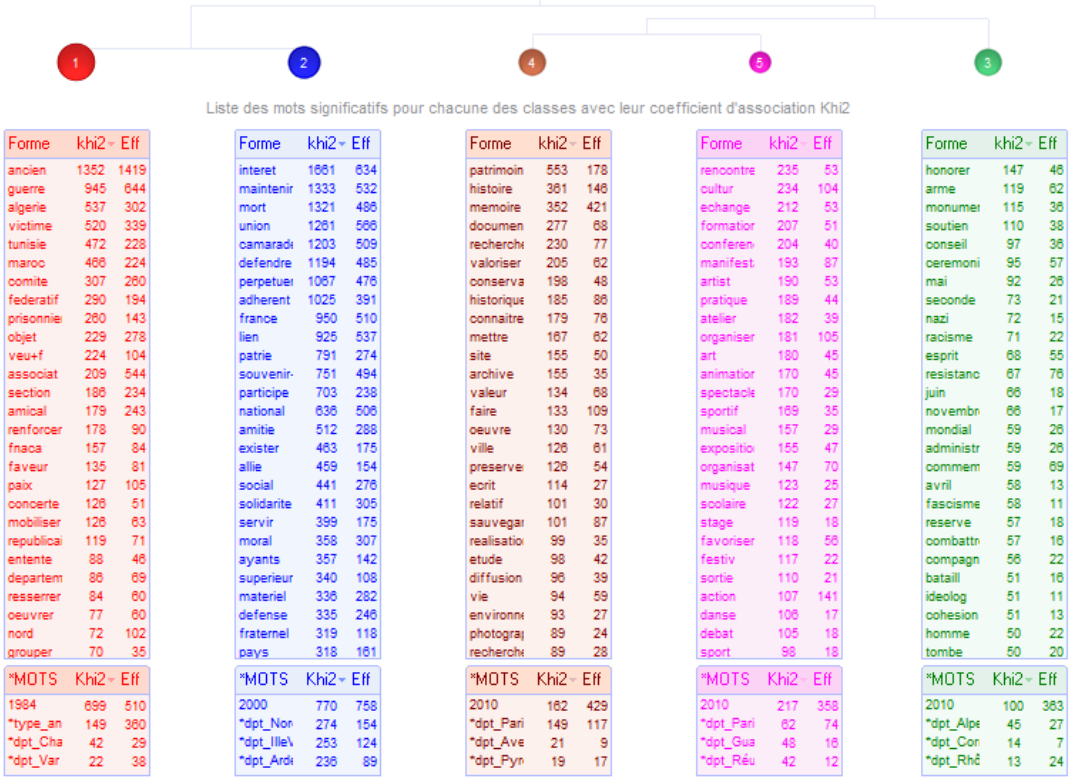


Figure 8 – Exemple de CDH du corpus global

[95] Concrètement, cette méthode établit un tableau lexical qui croise lignes (les UC, ie les individus) et colonnes (les formes réduites des mots, ie les variables) pour calculer la proximité lexicale entre UC, en fonction du nombre de leurs mots communs, cooccurrents. Les segments de texte d'une classe représentent ceux qui partagent le plus de mots communs.

[96] Une classe est donc caractérisée par des énoncés typiques, au sein desquels des formes sont surreprésentées. Cette "typicité" est mesurée par un khi2 d'appartenance, qui évalue la forte / faible appartenance d'une UC à une classe. Voir notamment RATINAUD P. et P. MARCHAND, « Application de la méthode ALCESTE à de «☐gros☐» corpus et stabilité des «☐mondes lexicaux☐»☐: analyse du «☐CableGate☐» avec IRaMuTeQ », *Actes des 11e Journées internationales d'Analyse statistique des Données Textuelles. JADT 2012*, 2012, p. 835‑844.

Les classes permettent donc de saisir la dynamique, la multiplicité discursive à l'œuvre dans le corpus. C'est pourquoi elles sont définies en elles-mêmes et par contraste aux autres classes (par les énoncés qui les spécifient). De cette méthode émerge la possibilité pour un même locuteur (ou un même document) d'être caractérisé par plusieurs types de discours : Alceste n'opère pas de typologie des individus mais repère la multiplicité des discours pouvant être exprimée au sein d'un même document ainsi que la similarité de discours pouvant être partagée par plusieurs documents.

**** *ID_92 *date_2010 *dpt_AlpesM *type_ann_1
association colloise de culture et partage recenser, rechercher , enregistrer tout ce qui concerne l'histoire, les traditions , les souvenirs ,le patrimoine concernant les villes et villages de la région, de les diffuser par tous moyens traditionnels ou multimédias lors des réunions ou manifestations diverses ayant trait au but de l'association.

**** *ID_93 *date_2010 *dpt_AlpesM *type_ann_1
association des rescapes et victimes du 5 juillet 1962 a oran sauvagarde de la mémoire du massacre perpétré le 5 juillet 1962 à oran ; organisation et manifestations sur sol national de commémoration du dit évènement ; installation de succursales sur sol national ; réalisation, diffusion, exploitation de tout matériel audio et visuel connu ou à venir, nécessaire à diffusion de ses messages ; acquisition de biens immobiliers aux fins d'installations de différents bureaux.

**** *ID_94 *date_2010 *dpt_AlpesM *type_ann_1
association de philipp friedrich mader soutenir et promouvoir des activités diverses culturelles auprès des protestants germanophones de nice et des alentours, et notamment : entretenir le souvenir du pasteur philipp friedrich mader et de son oeuvre ; cultiver la tradition et la vie culturelle allemande, voire la culture protestante, notamment dans son expression en france ; permettre à des français de découvrir la culture allemande, voire la culture protestante d'allemagne ; développer et diffuser des activités d'échanges internationaux entre les protestants français et les protestants allemands pour une compréhension mutuelle ; promouvoir les rencontres entre les

Figure 9 – Corpus « Déclarations d'associations au JO » : ici, 3 UCI

individu : 92 ****** *ID_92 *date_2010 *dpt_AlpesM *type_ann_1**
association colloise de culture et partage recenser, rechercher, enregistrer tout ce-qui concerne l' histoire, les traditions, les souvenirs, le patrimoine concernant les villes et villages de la region, de les diffuser par tous moyens traditionnels ou multimedias lors-des reunions ou manifestations diverses ayant trait au but de l' association.
individu : 93 ****** *ID_93 *date_2010 *dpt_AlpesM *type_ann_1**
association des rescapes et victimes du 5 juillet 1962 a oran sauvagarde de la memoire du massacre perpetre le 5 juillet 1962 a oran; organisation et manifestations sur sol national de commemoration du dit evenement; installation de succursales sur sol national; realisation, diffusion, exploitation de tout materiel audio et visuel connu ou a venir, necessaire a diffusion de ses messages; acquisition de biens immobiliers aux fins d' installations de differents bureaux.
individu : 94 ****** *ID_94 *date_2010 *dpt_AlpesM *type_ann_1**
association de philipp friedrich mader soutenir et promouvoir des activites diverses culturelles aupres des protestants germanophones de nice et des alentours, et notamment: entretenir le souvenir du pasteur philipp friedrich mader et de son oeuvre; cultiver la tradition et la vie culturelle allemande, voire la culture protestante, notamment dans son expression en france; permettre a des francais de decouvrir la culture allemande, voire la culture protestante d' allemagne; developper et diffuser des activites d' echanges internationaux entre les protestants francais et les protestants allemands pour une comprehension

Figure 10 – Corpus « Déclarations d'associations au JO » : distinction de 5 UCE (appartenant à 3 classes), issues des 3 UCI précédentes

4.4.4 Analyse discursive fine

Grâce aux dictionnaires intégrés, Alceste peut produire une analyse fine du vocabulaire d'un corpus. Les mots pleins lemmatisés sont en effet distingués au moyen de marqueurs grammaticaux et discursifs plus fins que ceux d'autres logiciels (tels Calliope). Ainsi, il est possible de différencier les mots pleins en noms, verbes et verbes modaux, adjectifs et

adverbes, nombres, entités nommées (prénoms, lieux / pays, mois / jour), couleurs et famille. S'ils n'entrent pas dans le processus d'élaboration des classes, les mots-outils peuvent néanmoins être sollicités lors de l'analyse des résultats pour spécifier davantage une classe de discours. Alceste repère ainsi divers marqueurs (modalisation, relations temporelle et spatiale, personne, intensité, relation discursive), auxiliaires être et avoir, démonstratifs, indéfinis et relatifs, qui sont autant de signes des types de discours.

A noter que cette analyse fine peut être précisée par les autres méthodes statistiques implémentées au sein d'Alceste. Une classification ascendante hiérarchique est ainsi proposée : si elle ne constitue pas la méthode employée pour élaborer les classes, elle fournit néanmoins un aide à la compréhension de la constitution de ces classes, en montrant comment les divers mots s'assemblent pour former une classe.

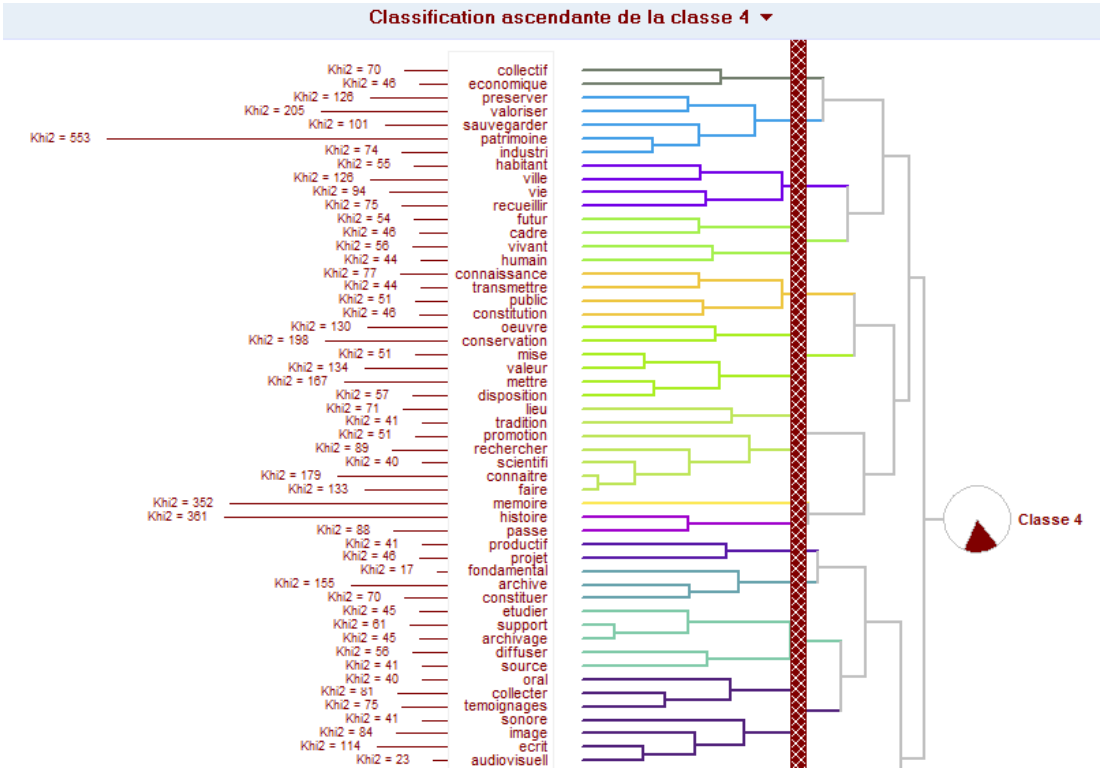


Figure 11 – Exemple de CAH – Les diverses classes composant la classe « Mémoire »

D'autres méthodes statistiques sont mobilisées au sein d'Alceste, telles que l'analyse factorielle de correspondances (AFC), qui permet de montrer le positionnement (opposition ou proximité) de classes selon des axes de valeurs.
Alceste propose également tout un éventail de résultats graphiques (concordancier, liste de segments répétés, réseaux de formes, courbes du vocabulaire, cartographique des unités textuelles, histogrammes, etc.) qui permettent d'affiner l'analyse.
Une analyse croisée est possible par l'ajout de variables illustratives documentant chaque « individu » (dans notre cas, les déclarations d'associations). Ces variables d'analyser la répartition des types de discours en fonction de leurs modalités : une date en vue d'une analyse chronologique des types de discours, un département pour une analyse de leur distribution spatiale, une catégorie socio-professionnelle, le type d'annonce des déclarations (création / modification), etc.

4.4.5 Alceste, une méthode d'accompagnement à la lecture

Dans la mesure où Alceste ne prétend pas révéler les thématiques d'un corpus, ce que disent les textes mais dévoiler les types de discours qui s'y entremêlent, il se veut davantage

une aide à la lecture, un premier décryptage de types de discours qu'il revient à l'analyste de vérifier, d'interpréter. Reinert nomme cette fonction exploratoire d'un corpus, une « lecture flottante », en ce qu'elle permet « l'imprégnation progressive des thèmes d'un discours et de leur discrimination dans un corpus donné [59, REINERT] ». Il s'agit non pas de mettre en lumière une des groupes de mots, isolés, mais de déceler les relations qui caractérisent une pratique discursive. La multitude des voix qui traversent un corpus textuel peuvent être invisibles à une lecture linéaire, humaine. C'est pourquoi le repérage de classes discursives contrastées peut aider l'analyste à identifier ces voix, ces postures énonciatives aux logiques et aux vocabulaires distincts. Alceste propose de cette façon, non pas l'exhibition de thématiques figées mais l'indice de possibles actes d'énonciation, dont la trace est statistiquement repérée au moyen de la cooccurrence de mots pleins appartenant à un même segment. La distribution lexicale devient alors une marque de l'activité énonciative dans sa « capacité d'animer rythmiquement des contenus ». « De même qu'un tableau d'Arcimboldo peut être objectivé différemment selon les perspectives sous lesquelles il se présente pour l'observateur, de même un texte est multiforme, matérialité traversée par des logiques possibles, dont l'orientation d'une lecture peut révéler la singularité. [59, REINERT]»

4.4.6 Mode opératoire d'Alceste

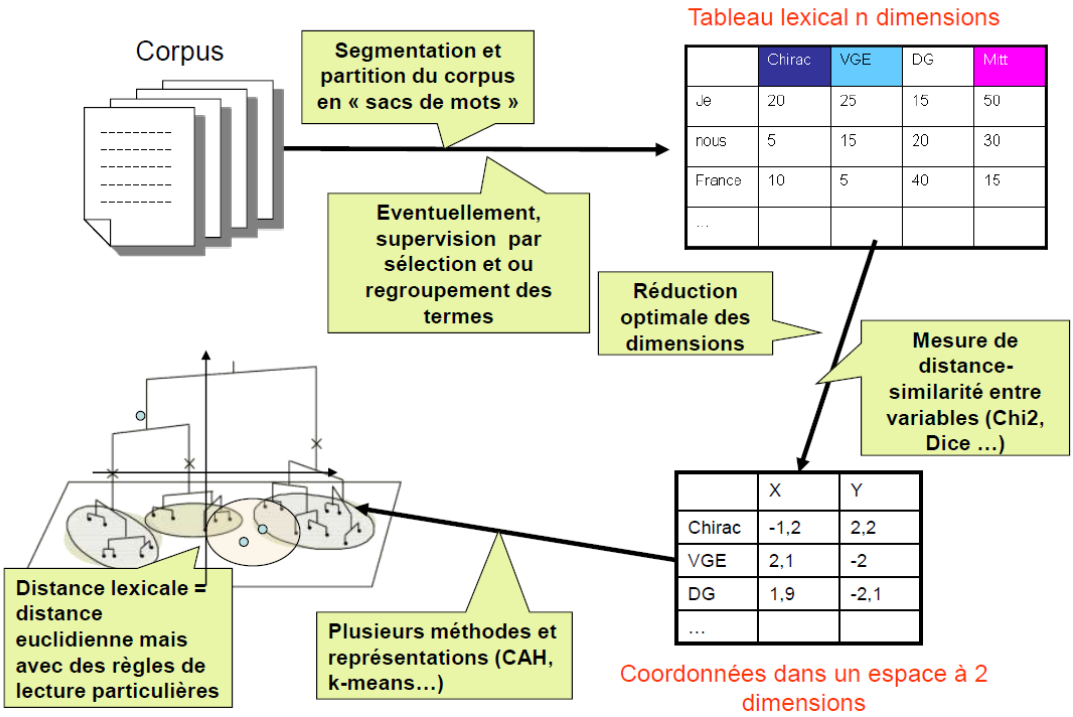


Figure 12 – Mode opératoire d'Alceste[97]

[97] « Panorama critique des analyses textuelles informatisées en SHS - Academia.edu », *op. cit.*

4.5 Tableau comparatif des modes de fonctionnement de Calliope et d'Alceste

En conclusion, nous présentons ci-dessous un tableau synthétique, résumant les méthodes sur lesquelles se fondent Calliope et d'Alceste, ainsi que leurs fonctionnalités (les éléments marqués d'un astérisque seront développés ultérieurement, dans partie « Méthodologie » et la partie « Présentation des résultats ») :

	Calliope	**Alceste**
Modes de traitements statistiques et fonctionnalités		
Méthode de classification*	Méthode des mots associés / Classification ascendante hiérarchique (CAH)	Classification descendante hiérarchique (CDH) + CAH en complément
Détection automatique de classes	Oui	Oui
Analyse factorielle des correspondances	Non	Oui, en complément
Intervention / supervision humaine lors du processus de classification*	Oui, validation du lexique d'indexation	Non
Spécification de marqueurs grammaticaux et discursifs* (outils linguistiques embarqués)	Faible	Fine
Identification entité nommées	Faible (noms propres)	Oui, fine
Modification possible des dictionnaires	Oui	Oui
Lemmatisation (et modification possible) *	Oui	Oui
Croisement avec des variables non textuelles*	Oui	Oui
Extraction de sous-corpus et tri croisé	Oui, non complet	Oui
Gestion et présentation des résultats*		
Retour aux textes initiaux	Oui	Oui
Mode d'interrogation des textes	Opérateurs booléens (ET / OU)	Requête simple (absence d'opérateurs booléens)
Représentations graphiques des résultats statistiques	Oui, multiples + (diagramme stratégique, clusters avec liens internes et externes, représentations chronologiques)	Oui, multiples ++ (CDH, CAH, AFC, clusters, histogrammes, disques, listes unités textuelles, segments répétés, etc.)
Navigation entre représentations graphiques	Oui	Oui
Accès à concordancier, segments répétés*	Non	Oui
Lisibilité des résultats	Oui	Parfois difficile (AFC)
Qualité de l'interface graphique	Moyenne	Moyenne
Edition de rapports, d'aides à	Non	Oui, complet

l'interprétation		
Coût d'entrée		
Coût du logiciel	Gratuit	Licence payante
Coût d'apprentissage (prise en main)	Moyen	Elevé

Deuxième partie
Méthodologie
et approche comparative des outils d'analyse de données textuelles

1 Description de la source

Cette deuxième partie va exposer les principes méthodologiques qui ont été appliqués dans notre projet d'analyse de données textuelles du lexique mémoriel. Elle débutera par une présentation de la source de données sélectionnée (déclarations d'associations aux *Journal officiel*) et des différentes phases de constitution du corpus. Elle sera suivie par l'exposé, dans une perspective comparative, des différents prétraitements et traitements qui ont été effectués sur les données et sera conclue par une analyse critique de la méthode employée et des limites tant internes (accessibilité problématique des données, par exemple) qu'externes (durée courte du stage) qui ont affectées le projet.

Nous allons désormais décrire la méthodologie qui a présidé à l'élaboration du corpus. Loin d'être un cheminement linéaire se déployant en étapes préalablement et définitivement fixées, la constitution du corpus suit une démarche nécessairement itérative, récursive et progressive. Cette démarche, constituée d'allers-retours, ne détermine les conditions initiales du projet que pour mieux les modifier au fur et à mesure de l'avancée du projet, en fonction des résultats issus des choix effectués à chaque étape. La boucle se substitue à la ligne.
L'itérativité de la démarche se justifie particulièrement dans le cadre d'un projet exploratoire, devant composer avec des contraintes tant externes (durée du stage, degré d'accessibilité et de qualité des données, prétraitements requis par les outils d'analyse textuelle utilisés, etc.) qu'internes (sélection des déclarations selon leur pertinence pour le projet de recherche).
Le dimensionnement de notre projet – comme tout projet s'inscrivant dans le domaine des humanités numériques – est donc délimité par les ressources qui lui sont affectées, dont les limites temporelles constitue un facteur déterminant. Ces limites conditionnent la sphère du possible et définissent le périmètre des résultats obtenus. Ceux-ci ne constituent qu'une étape temporaire à partir de laquelle se fonder pour affiner l'exploration des données.
Les choix retenus dans cette étude découlent d'une première appréhension quantitative des fichiers à notre disposition et de nombreux échanges et discussions avec les personnes impliquées dans le projet (au premier chef desquelles, la chercheuse, Sarah Gensburger), de manière à ce que la sélection opérée entre les différentes options qui se présentaient au terme d'une étape satisfasse les orientations scientifiques du projet et soit adaptée à sa problématique.
Précisons, enfin, que le matériau textuel étudié dans le cadre de ce projet est le même dans le cas des deux outils d'analyse de données textuelles sélectionnés, Calliope et Alceste. Cette identité de corpus sert ainsi une double finalité : permettre non seulement de souligner la spécificité de chaque logiciel mais aussi de stimuler les interrogations de la recherche, par un éclairage différent et complémentaire du même matériau.
Nous présenterons brièvement la source de données sélectionnée (les déclarations d'associations au *Journal officiel*) ainsi que les différentes étapes de la constitution du corpus.

1.1 Déclarations d'associations au *Journal officiel*

Le corpus choisi pour nous permettre d'étudier la distribution et l'évolution des thèmes mémoriels des acteurs sociaux est constitué de déclarations d'associations[98] publiées au

[98] Le choix de ce type de source diffère sensiblement des matériaux textuels habituellement soumis aux outils d'analyse textuelle, qui privilégient plutôt les « échantillons d'entretiens de recherche, les dossiers de presse, les ensembles de documents d'archives, ou toute autre composition de matériaux rassemblés ou produits dans le cadre d'une enquête, et donc constitués en fonction d'un questionnement sociologique », *in* DEMAZIÈRE D., C. BROSSAUD, P. TRABAL, et K.M. METER (VAN) (dir.), *Analyses textuelles en sociologie Logiciels, méthodes, usages Karl (dir.), op. cit.*

Journal officiel « Lois et décrets » d'abord, puis au *Journal officiel « Associations et Fondations d'entreprise »* (JOAFE). Ces publications officielles sont éditées par la Direction de l'information légale et administrative (DILA), service du premier ministre.
Ce bulletin publie les avis de création, de modification et de dissolution des associations régies par la loi du 1er juillet 1901, des associations syndicales de propriétaires, des fondations d'entreprise et des fonds de dotation.
Dans le cadre de notre travail, nous nous intéresserons exclusivement aux déclarations d'associations, comme représentative d'une partie des acteurs sociaux.

Présentation du *Journal officiel* (JO)
Le *Journal officiel* de la République française (JORF ou simplement JO[99]) est la publication officielle éditée par l'État français, dans laquelle sont consignés tous les évènements législatifs (lois), réglementaires (arrêtés, décrets), déclarations officielles et publications légales. Il est sous-titré « Lois et décrets » ou « Édition des documents administratifs » selon la nature des textes qu'il contient.
Le JO est une publication quotidienne dont la date de parution conditionne la date à laquelle le texte produit des effets juridiques (un texte à valeur juridique doit être connu, donc rendu public, pour pouvoir être applicable).
Le 26 mars 2015, le premier ministre, M. Valls, a annoncé la disparition de la version papier pour la fin de l'année 2016. Le format électronique a donc valeur juridique.

***Journal officiel* « Associations et Fondations d'entreprise » (JOAFE)**
Créée en 1985, le *Journal officiel « Associations et Fondations d'entreprise »* (JOAFE) prend la forme d'un supplément hebdomadaire de l'édition « *Lois et décrets »* du *Journal officiel*, se substituant à la rubrique spéciale des annonces des associations que contenait auparavant le JO.
Précisons d'emblée que ce changement de support et de périodicité a entraîné des modifications, dans le cadre de notre travail, aux niveaux de la quantité de fichiers fournis (d'une publication quotidienne, le support devient hebdomadaire, réduisant ainsi drastiquement le nombre de fichiers à organiser) et de la structuration des déclarations – modifications qui ne seront pas sans conséquence sur le travail de normalisation des données que nous décrirons plus tard.

Associations, capacité juridique et publicité légale
La *Loi du 1er juillet 1901*, relative au contrat d'association, constitue la base du droit des associations en France[100]. Ce texte reposant sur le principe de la liberté contractuelle, le contrat d'association peut demeurer un acte purement privé (article 2[101]). Cependant, en vertu de l'article 5[102], les dirigeants, s'ils le souhaitent, peuvent conférer la personnalité morale à l'association en en faisant reconnaître l'existence par les pouvoirs publics. Ce statut

---

[99] Pour des questions de simplification, nous désignerons notre source, dans la suite de ce mémoire, par le terme *Journal officiel* (ou JO).
[100] « L'association est la convention par laquelle deux ou plusieurs personnes mettent en commun, d'une façon permanente, leurs connaissances ou leur activité dans un but autre que de partager des bénéfices. Elle est régie, quant à sa validité, par les principes généraux du droit applicables aux contrats et obligations. », », *Loi du 1er juillet 1901 relative au contrat d'association*, Titre 1, Article 1
[101] « Les associations de personnes pourront se former librement sans autorisation ni déclaration préalable, mais elles ne jouiront de la capacité juridique que si elles se sont conformées aux dispositions de l'article 5. », *Ibid.*, Titre 1, Article 2
[102] « Toute association qui voudra obtenir la capacité juridique prévue par l'article 6 devra être rendue publique par les soins de ses fondateurs. La déclaration préalable en sera faite au représentant de l'Etat dans le département où l'association aura son siège social. Elle fera connaître le titre et l'objet de l'association, le siège de ses établissements et les noms, professions et domiciles et nationalités de ceux qui, à un titre quelconque, sont chargés de son administration. Un exemplaire des statuts est joint à la déclaration. Il sera donné récépissé de celle-ci dans le délai de cinq jours. L'association n'est rendue publique que par une insertion au *Journal officiel*, sur production de ce récépissé. Les associations sont tenues de faire connaître, dans les trois mois, tous les changements survenus dans leur administration, ainsi que toutes les modifications apportées à leurs statuts. », *Ibid.*, Titre 1, Article 5

confère à l'association la capacité juridique d'agir en son nom propre pour le compte de ses membres dans l'accomplissement de son objet.
L'acquisition de la personnalité juridique est soumise à une double condition de publicité :
- La déclaration de l'association à l'autorité préfectorale (préfecture du département ou sous-préfecture de l'arrondissement du siège social de l'association),
- L'insertion au *Journal officiel* d'un extrait de cette déclaration.
La possibilité est offerte aux membres de modifier l'objet, les règles de fonctionnement, le siège social d'une association, etc., en publiant des avis de modification et, le cas échéant, de dissolution de l'association.

Le statut officiel de la déclaration explique, en outre, le caractère descriptif et formaté de son contenu (le nombre et le type de champs sont contraints), qui le rend particulièrement adapté aux outils d'analyse de données textuelles orientés « lexique » que nous avons sélectionnés.

Couverture géographique des déclarations au JO
Les publications du JO couvrent l'intégralité du territoire national, c'est-à-dire l'ensemble des départements français. Elles ne concernent donc pas les associations dont le siège social est situé dans les autres territoires de la République française que sont les collectivités d'outre-mer (Nouvelle Calédonie, les pays d'outre-mer (Polynésie Française et Wallis et Futuna), ainsi qu'un territoire à statut particulier (Terres Australes et Antarctiques). Ces territoires étant soumis à une législation particulière, leurs annonces ne sont pas publiées au *Journal officiel « Associations et fondations d'entreprise »* mais bénéficient d'une publicité légale locale.

Politique d'ouverture des données publiques (*Open Data*[103])
La DILA, dont l'une des principales missions est d'assurer la diffusion des données dont la publication au JO et dans les bulletins d'annonces légales est obligatoire, remplit par ailleurs une autre fonction, complémentaire. Elle se doit en effet de faciliter, par la mise en place de services et de moyens technologiques, l'accès des citoyens à l'information législative et documentaire. Cette fonction implique une attention portée à l'évolution des besoins et des techniques, en particulier dans le domaine de la dématérialisation de l'information.
S'inscrivant donc « naturellement » dans la nouvelle politique d'ouverture des données publiques, la DILA favorise la démocratisation de l'accès à l'information qu'elle diffuse. Elle rend ainsi possible l'accès aux publications du JO, via des fichiers XML (pour les déclarations à partir de 1997). Pour les publications antérieures à 1985 (1939-1984), les fichiers papier ont été scannés par la DILA. Pour la période 1985-1996, les déclarations sont archivées sous forme de microfiches à la BnF – ce qui explique que, dans le cadre du projet de recherche, seules les déclarations des départements 75 et 38 (Isère) ont été pour l'instant traitées, car ce traitement est manuel.[104]
Outre le souci de transparence démocratique et de partage des informations détenues par la puissance publique, la mise en place de cette politique d'ouverture des données publiques impacte directement les objets et méthode de la recherche scientifique. Accéder à des jeux de données « inédits », pouvoir les manipuler élargit le champ des investigations scientifiques. Elle a de fait rendu possible le projet de recherche dans lequel nous nous inscrivons.

La structuration des déclarations : la DTD

---

[103] Pour une mise en perspective de l'ouverture des données et la fabrication des « données brutes », voir DENIS J. et S. GOËTA, « La fabrique des données brutes. Le travail en coulisses de l'open data », Paris, 2013.
[104] Cet accès aux publications officielles peut se révéler plus complexe que prévu. Voir, à cet effet, l'article d'Emilien Ruiz, « Accéder aux numérisations du *Journal officiel* de la République française, de 1871 à nos jours », sur le site La boîte à outils des historiens : http://www.boiteaoutils.info/2013/01/acceder-aux-numerisations-du-journal.

Les déclarations publiées au JO sont structurées en différents champs, enregistrant des informations de nature diverse. L'ensemble de ces catégories est défini dans un document, *Documentation technique du JOAFE*[105], publié par la DILA et qui contient la DTD (Document Type Definition[106]) des fichiers XML des déclarations.

Cette DTD comprend les éléments suivants :

- **Elément « Parution »** : contient une date et un numéro de parution
- **Elément « Annonce »** : correspond à la publication d'un avis, avec l'ensemble des informations caractérisant une association (numéro de l'annonce, lieu de déclaration, titre, objet, siège social, etc.)
- **Elément « Type »** : identifie le type d'avis (création, modification, dissolution) au moyen d'un code.
  Seuls les avis de création et de modification ont fait l'objet d'une analyse, les dissolutions ayant été écartées en raison de l'absence d'informations obligatoires les concernant[107].
- **Elément » Thème »** : indexe chaque déclaration à un ou plusieurs thèmes concernant leur domaine d'activité, comme la défense des droits fondamentaux, les pratiques d'activités artistiques ou encore, l'aide à l'emploi et développement local.

C'est sur l'élément « Annonce » que nous appliquerons les techniques lexicométriques, car il contient l'intégralité du matériau textuel des déclarations nous intéressant (Titre et Objet). Dans le cadre d'un prolongement de notre travail (hors stage), une analyse de la distribution du lexique par type d'annonce (création / modification d'une déclaration) sera également effectuée, afin de voir notamment si les associations déjà existantes modifient (ou non) leur contenu pour intégrer des termes mémoriels, ou si de nouvelles associations se créent avec ce vocabulaire (ou non). Cela permettra d'affiner l'analyse chronologique du lexique des associations.
Par ailleurs, cette approche par « contenu » devrait être complétée par une étude des données biographiques des membres des associations, et de leur trajectoire, dans le cadre d'une analyse prosopographique, second volet du projet de recherche.

1.2 Autres sources de données non traitées dans le cadre de cette étude

Il existe de nombreuses autres sources d'information sur les associations en France, qui auraient également pu faire l'objet d'un traitement par des outils de textométrie
Nous en décrivons quelques-unes ci-dessous, qui pourraient, à l'avenir être étudiées dans le cadre d'un prolongement du projet de recherche.

- **Microfiches des déclarations du JO « Associations » archivées à la BnF (1985-1996)**
  Ces fichiers correspondent à une partie de notre source (JO Associations), non disponibles en XML, et donc traitables dans le cadre de notre stage.

---

[105] Cette structuration ne concerne que les fichiers publiés depuis 1997.
[106] DTD : « document permettant de décrire un modèle de document SGML ou XML. Le modèle est décrit comme une grammaire de classe de documents : *grammaire* parce qu'il décrit la position des termes les uns par rapport aux autres, *classe* parce qu'il forme une généralisation d'un domaine particulier, et *document* parce qu'on peut former avec un texte complet. Au niveau de la structure logique, une DTD indique les noms des éléments pouvant apparaître et leur contenu, c'est-à-dire les sous-éléments et les attributs. En dehors des attributs, le contenu est spécifié en indiquant le nom, l'ordre et le nombre d'occurrences autorisées des sous-éléments. L'ensemble constitue la définition des hiérarchies valides d'éléments et de texte. », https://fr.wikipedia.org/wiki/Document_type_definition. Voir en annexe la DTD détaillée du JOAFE (p. 223).
[107] Cet aspect sera détaillée dans la partie suivante « Méthodologie de constitution du corpus ».

Pour des questions de temps, seuls les fichiers de Paris et de l'Isère ont été jusqu'ici traités manuellement et analysés. Cette analyse repose sur une sélection de fichiers, après filtrage des champs « Titre » et « Objet » selon 5 mots-clés (mémoire / commémoration / célébration / souvenir / anciens combattants).
Couverture temporelle : 1985-1996
Couverture géographique : Territoire national
Formats de fichiers : Microfiches
Source : DILA (JO)
A noter : les fichiers de Paris sont caractérisés par les termes « mémoire » et « monde combattant » tandis que ceux de l'Isère le sont par « mémoire » et « droits de l'homme ».
Dans le cadre d'un futur projet, seront traités et analysés les fichiers des départements suivants : Bouches du Rhône, Cantal, Isère, Nantes, Rhône, Paris, Hauts de Seine, Seine-Saint-Denis, Martinique. Ces départements ont été choisis en raison de caractéristiques multiples, permettant d'établir un échantillon varié des départements français (départements fortement urbanisé ou à dominante rurale, diversité des catégories socio-professionnelles de leur population, liens multiples aux phénomènes mémoriels comme le projet de création d'un lieu mémoriel à Villeurbanne, mémoire de l'esclavage, etc.).

- **Fichiers SIMPA (**Système d'Information Multi-services des Partenaires Associatifs[108]) **des associations en relation avec la Mairie de Paris**
  Ce système, mis en place par la Mairie de Paris, s'inscrit dans la même démarche que le répertoire RNA de simplification et de dématérialisation des démarches administratives. Il s'adresse aux associations loi 1901 qui sont en relation avec la Ville de Paris, domiciliées à Paris ou non, mais exerçant tout ou partie de leurs activités sur le territoire parisien.
  Indexation : lors de l'enregistrement de sa déclaration, l'association peut sélectionner parmi un vocabulaire contrôlé de 22 secteurs d'activité, celui (ou ceux, dans la limite de trois) qui la représente(nt) le mieux. Parmi ces descripteurs, se trouve le terme « Mémoire »[109].
  Couverture temporelle : 2008-2013
  Couverture géographique : Paris
  Source : Mairie de Paris
  Les fichiers de la Mairie de Paris ont été traités avec Calliope par Mathilde de Saint-Léger. Nous avons utilisé la terminologie validée, par Sarah Gensburger, comme base pour la validation du lexique extrait des fichiers du JO.

D'autres fichiers du même type, mis en place et gérés par les municipalités françaises, existent. Sarah Gensburger, dans l'optique de futurs projets de recherche complétant celui en cours, en a fait la demande auprès des 40 plus grandes villes de France.

- **Réseau Mémoire et histoire en Ile-de-France**[110], qui organise tous les 2 ans, le « Printemps de la mémoire ».
  Ce réseau, créé en 2010, regroupe les « associations qui mènent – avec une démarche transversale et transdisciplinaire – un travail de mutualisation et de réflexion sur le champ de la mémoire et de l'histoire sociale et culturelle sur les grandes thématiques des questions urbaines, des migrations, du monde du travail ».

---

[108] Voir le site : http://blogs.paris.fr/simpa/simpa.
[109] Contrairement aux catégories de l'élément « Thème » du JO, qui ne contiennent pas le terme « Mémoire ». Autre différence distinguant les deux systèmes d'indexation, l'affectation de descripteurs d'activité est un processus auto-déclaratif dans les fichiers SIMPA, tandis qu'il est le fait des autorités préfectorales et du service éditeur de la DILA, dans le cas des JO. Cette particularité constitue l'une des raisons pour lesquelles l'élément « Thème » n'a pas été retenu dans le cadre de l'analyse lexicale des fichiers du JO – comme nous le verrons plus loin.
[110] Voir le site : http://www.memoires-histoires.org.

Il répertorie les associations d'Ile-de-France membres du réseau et/ou qui ont participé à au moins une édition du « Printemps de la mémoire » (biennale).
> Métadonnées : nom sans objet de l'association
Couverture temporelle : 2010-
Zone géographique : Ile-de-France
Source : Réseau Mémoire et histoire en IDF

2 Méthodologie de constitution du corpus

Comme nous l'avons montré précédemment (partie I), la constitution du corpus[111] à analyser constitue une phase essentielle, bien que délicate, dans toute démarche d'analyse textuelle et, de façon plus générale, dans tout traitement de corpus – que ce soit de façon manuelle ou au moyen d'outils informatiques. Loin d'être anodine, cette étape requiert vigilance et questionnement. Les choix, notamment méthodologiques, présidant au rassemblement de documents au sein d'un corpus conditionnent en effet les résultats et les analyses futurs. C'est la raison pour laquelle chaque étape du processus doit être, afin d'en assurer la pertinence scientifique, l'objet de questionnements et de discussions avec les chercheurs.
Les paragraphes qui suivent pourront paraître bien sommaires, évidents, voire naïfs ou erronés aux chercheurs, notamment linguistes, qui traitent quotidiennement des difficultés de constitution de corpus. Ils se veulent simplement la trace de notre étonnement, et parfois de notre désarroi, face à des situations impliquant la suppression d'actions considérées comme n'étant plus pertinentes et un retour à l'étape antérieure. Ces remarques nous semblent néanmoins importantes à formuler dans le contexte de la textométrie, dans la mesure où l'apparente facilité et fiabilité qu'offrent les outils s'appuyant sur des principes statistiques, quantitatifs, tend à minorer ou dissimuler la part de réflexion nécessaire à introduire en amont et en aval des traitements proprement dits. Ces outils ne sauraient se réduire à de simples « presse-boutons », selon l'expression favorite de Mathilde de Saint-Léger, mais s'inscrivent au contraire dans un processus où le quantitatif se nourrit de qualitatif, et inversement.

La constitution d'un corpus doit être guidée par certaines précautions méthodologiques qu'il s'agit de garder à l'esprit, afin d'en limiter les risques de biais[112] :
- Quelle est la fonction du corpus à construire ? Est-ce de valider des hypothèses de recherche ou de cerner des « phénomènes » permettant de mieux délimiter l'objet de recherche, d'une manière exploratoire ?
- De quelle « réalité » le corpus doit-il être représentatif ? Quel phénomène s'agit-il d'appréhender par son traitement ? Quels documents prendre en compte ? Lesquels exclure du corpus ? Quel découpage effectuer des données ? Quels réagencement sont possibles ?
- Comment délimiter sa couverture temporelle et géographique ?
- Les données disponibles sont-elles hétérogènes ? Si oui, quelles méthodes adopter pour les normaliser ?
- Quel est le dimensionnement et la faisabilité du projet ? Comment calibrer chaque étape, en termes de ressources ?

Nous essaierons de montrer, dans les paragraphes qui suivent, comment nous avons tenté de répondre à ces questions.

2.1 Adéquation de la source à la méthode d'analyse de données textuelles

Les sources potentielles de documents concernant les associations sont, comme nous l'avons vu, variées (les fichiers des déclarations au JO, les fichiers SIMPA de la Mairie de Paris, les données du Réseau mémoire et histoire en Ile-de-France, etc.). Dans le cadre de notre

[111] Pour une présentation des difficultés inhérentes à la constitution d'un corpus, voir la première partie de ce mémoire.
[112] Ces précautions méthodologiques sont inspirées de DALBERA J.-P., « Le corpus entre données, analyse et théorie », *op. cit.* et de OLLIVIER G., « Panorama critique des analyses textuelles informatisées en SHS - Academia.edu », *op. cit.*

étude, notre choix s'est porté sur les déclarations des associations au JO pour différentes raisons : disponibilité – relative, comme nous le verrons un peu plus loin –, des fichiers, type d'informations contenues (notamment, titre et objet de la déclaration, ainsi que les modifications apportées), amplitude de la période temporelle (de 1947 à nos jours) et de l'aire géographique (ensemble des départements français) couvertes.
Le choix de cette source de données constitue une première orientation des résultats. Le lexique des déclarations du JO diffère, bien évidemment, de celui des fichiers SIMPA. Quels que soient les résultats produits, ils n'auront de validité et de pertinence qu'au regard de leur source.

La richesse des déclarations au JO rend possible le traitement et l'analyse de différentes caractéristiques du corpus, par son partitionnement du corpus en sous-corpus ou par l'analyse croisée selon des variables qualitatives (comme le département) : analyse de la répartition temporelle et spatiale du lexique, analyse des modifications des déclarations (modification de l'intitulé de l'association, de son objet et dans le cas d'une analyse prosopographique, de son siège social et des membres composant son bureau).

Dans l'approche comparative qui est la nôtre, nous avons constitué un corpus unique nous permettant d'évaluer et de confronter les méthodes et les résultats des deux outils d'analyse textuelle sélectionnés, Calliope et Alceste. Et de pourvoir en dégager, *in fine*, les similarités, divergences et éventuelles complémentarités.

2.1.1 Critères de choix du JO

Nous allons désormais détailler les critères justifiant le choix du JO comme source de notre corpus. Cette source se caractérise les propriétés suivantes :
- auto-déclarativité des producteurs de documents,
- large couverture temporelle et géographique des déclarations,
- formatage de la structure et du contenu textuel,
- adéquation du matériau textuel aux outils de lexicométrie sélectionnés.

2.1.1.1 Auto-déclarativité : proximité de la source

L'intérêt premier, dans le cadre d'une étude du vocabulaire des associations, de sélectionner les déclarations des associations au *Journal officiel* est d'avoir accès à des documents rédigés par les membres de l'association elle-même, et donc de se situer au plus près des « producteurs » – même si cette affirmation doit être aussitôt nuancée par les contraintes qu'impose la rédaction de tout acte administratif et officiel, qui implique un certain formatage du contenu textuel.
Les déclarations étant renseignées par les sociétaires de l'association (membres du bureau), le choix des termes les désignant et exprimant la finalité de l'association ne peut qu'être significatif pour qui veut tenter de comprendre la manière dont les associations se (re)présentent dans un document ayant valeur juridique, et donc destiné à être diffusé publiquement.
Les informations de certaines des autres sources mentionnées plus haut, comme le Réseau mémoire et histoire en Ile-de-France, sont moins pertinentes pour notre objet, dans la mesure où elles ne sont pas auto-déclaratives.

2.1.1.2 Large couverture temporelle et géographique

L'autre avantage du JO réside dans sa large couverture à la fois temporelle (pour la période contemporaine qui nous intéresse, de 1939 à nos jours) et géographique (territoire national). Cette propriété rend possible des analyses croisées selon plusieurs variables, de plus ou moins fine granularité. Une analyse d'envergure, englobant l'ensemble des fichiers ou, au contraire, un zoom sur une période historique et/ou une zone géographique sont identiquement envisageables.

Les autres sources à notre disposition sur les associations ne possèdent pas, à des degrés divers, une telle couverture.

2.1.1.3 Formatage des déclarations : homogénéité de la structure et du contenu

Le type de document enregistrant les déclarations au JO (formulaire) explique que celles-ci soient constituées d'éléments textuels formatés, dont le nombre et le type de champs sont définis. Le contenu a une fonction descriptive et sa taille limitée (objet), voire strictement délimitée (250 caractères maximum pour le titre). Ce formatage représente un atout pour le traitement statistique des données en ce qu'il contribue à homogénéiser le contenu et la structure des documents (en qualité et en quantité).
La petite dimension des éléments textuels nous a néanmoins obligée à concaténer certains champs, afin d'atteindre une taille pertinente d'un point de vue statistique. Ainsi, la petite taille du titre des associations, voire même de leur objet (plus de 94% des déclarations ont un contenu, titre et objet, compris entre 100 et 1000 caractères) n'a pas permis d'appliquer les techniques d'analyse textuelle sur ce seul champ. Le matériel textuel étudié dans le cadre de ce travail comprend donc toujours, conjointement, titre et objet de l'association[113].

2.1.1.4 Adéquation du matériau textuel aux outils d'analyse de contenu

Le type de contenu des déclarations est, par ailleurs, parfaitement adapté aux outils de lexicométrie retenus dans le cadre de cette étude. Ce contenu est en effet caractérisé par des textes à teneur descriptive, composé essentiellement de substantifs, d'adjectifs et de verbes à l'infinitif, qui se prêtent bien à l'analyse de contenu (« ce qui est dit »). Leur niveau de complexité langagière s'avère relativement peu complexe (absence d'interactions langagières, de procédés rhétoriques, etc.), contrairement à d'autres types de documents (comptes rendus d'entretiens, textes littéraires, documents épistolaires, etc.).
Le type de textes constituant notre corpus est ainsi parfaitement adapté aux méthodes implémentées au sein des outils d'analyse textuelle que nous avons sélectionnés :
- Calliope, tout d'abord, qui opère une réduction du lexique aux seuls mots pleins[114] lemmatisés[115]. Sa non prise en compte, dans l'analyse de la distribution lexicale, des marqueurs de l'énonciation que sont par exemple les déictiques (adverbes de lieu et de temps, pronoms possessifs et démonstratifs, etc.) – ces derniers organisant les relations spatiales et temporelles de l'énonciation[116], convient donc bien au type de contenu des déclarations.
Après une première analyse, nous avons par ailleurs décidé de ne pas conserver les verbes ni les adjectifs lors de l'extraction du lexique. En effet, les résultats se présentant sous la forme de clusters de termes, l'adjonction de verbes au sein de ces clusters ne représente pas de plus-value informationnelle. Tout au plus apprend-on que « souvenir » est fortement corrélé au verbe « perpétuer ». L'absence de segments répétés, l'importance accordée aux noms et aux thématiques qu'ils représentent, font de Calliope, à notre avis, un outil de « lexicométrie nominale ». Les adjectifs sont également été supprimés de l'analyse en raison de leur trop grande polysémie

---

[113] Nous verrons plus loin la limite que peut constituer cette concaténation des champs.
[114] Les mots pleins ou mots lexicaux désignent les mots dont la fonction sémantique est au moins aussi importante que la fonction syntaxique (noms, adjectifs qualificatifs, verbes et adverbes), tandis que les mots-outils caractérisent la catégorie de mots dont le rôle syntaxique est supérieur au rôle sémantique (articles, adjectifs non qualificatifs, pronoms, etc.).
[115] Comme définie en première partie, la « lemmatisation » est le processus linguistique qui consiste à regrouper les différentes formes fléchies (conjugaison, accord en nombre et en genre, etc.) d'un mot sous leur forme canonique ou lemme (comme la forme infinitive pour un verbe ou le masculin singulier pour un substantif). Ce processus ne doit pas être confondu avec celui de « racinisation », qui consiste à réduire, en supprimant les préfixes et les suffixes, les différentes formes d'un mot en son radical ou racine. Ainsi, le radical de « chercher » ou « chercheur » est « cherch », qui ne correspond pas à un mot réel de la langue, contrairement au lemme.
Nous avons vu en première partie que ce processus constitue une hypothèse forte quant à l'interprétation de ce qui est traité comme matériau textuel, qui n'est pas sans conséquence sur les résultats obtenus.
[116] Voir SARFATI G.-E., *Eléments d'analyse du discours*, Armand Colin., Paris, 1997.

- Alceste, de même, ne retient pour son analyse que les mots pleins lemmatisés, les mots-outils ne participant pas au processus d'élaboration des classes. Mais, contrairement à Calliope, sa catégorisation grammaticale est plus fine et il est possible d'étudier la distribution de ces catégories au sein des classes. A noter également la possibilité offerte d'accéder aux segments répétés, repérés au sein du corpus, qui nuance la prédominance attribuée aux mots pleins dans l'analyse. Ces segments peuvent en effet être des syntagmes nominaux et verbaux, bien sûr, mais aussi prépositionnels.

Cette adéquation du matériau textuel des déclarations aux postulats langagiers embarqués au sein des outils limite, en outre, les risques inhérents que peuvent, précisément, représenter ces postulats. Comme nous l'avons vu en première partie, les diverses opérations de réduction du lexique (lemmatisation, catégorisation grammaticale notamment), effectuées en vue de son traitement statistique, ne sont anodines et influencent les résultats. Le fait que le lexique du JO corresponde aux principes linguistiques et statistiques (primat du mot plein, importance de la fréquence relative, etc.) des outils minimise donc les risques de perte et de distorsion d'informations que ces processus peuvent faire subir au langage. Nous verrons néanmoins dans la troisième partie de ce mémoire que cette adéquation n'écarte pas tout risque potentiel.

2.1.1.5 Les déclarations du JO : une représentation (dé)limitée des acteurs sociaux

Avant de clore cette présentation, il faut noter que le fait de sélectionner les déclarations du JO comme source du corpus constitue une première réduction et orientation de l'objet initial de la recherche. Selon la méthode adoptée, selon la source choisie, le corpus retenu, non seulement le mode de (re)présentation des acteurs sociaux, et des associations en particulier, varie mais aussi les interprétations qui en seront issues. En cela, la sélection des déclarations au JO comme source de notre étude oriente nécessairement l'analyse qui pourra en être faite. L'utilisation d'outils d'analyse textuelle ne saurait en effet dispenser de resituer les discours analysés en lien à leur contexte de production et d'énonciation. Il ne faut donc pas oublier, comme le soulignent Demazière *et al.* (p. 179), que les textes collectés pour former corpus sont « une production sociale, engageant des acteurs ayant des caractéristiques biographiques spécifiques et poursuivant des stratégies au principe même de la production des matériaux. »

Loin de représenter l'ensemble des acteurs sociaux, les déclarations du JO ne retiennent de cette catégorie que les associations. D'autres membres composent la « société civile », tels que les médias, les organisations caritatives, les organisations non gouvernementales, les communautés religieuses, etc.

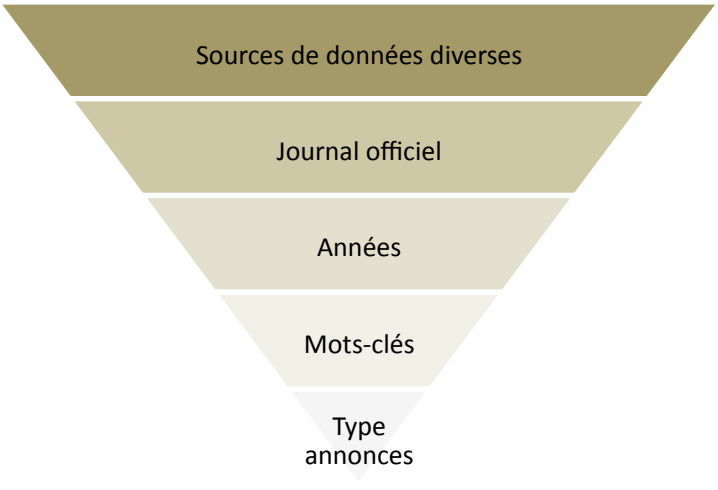
Le type même des documents sélectionnés pour former le corpus (les déclarations du JO) constitue, par ailleurs, une autre forme de réduction – ou une autre forme d'interprétation – de la représentation de la demande sociale. L'évolution du lexique propre aux associations aurait ainsi pu être appréhendée autrement, via l'analyse d'autres documents : étude de leurs activités et événements, de leurs supports de communication (internes et externes), des comptes rendus d'assemblée générale, des déclarations publiques et discours officiels, de leurs interventions dans la presse, de leurs réseaux sociaux, réalisation d'entretiens avec les membres dirigeants ou les simples adhérents, d'observations, etc. Enfin, les déclarations publiées au JO par la DILA ne représentent qu'un extrait des déclarations déposées auprès des services préfectoraux. En sont exclues, pour des raisons évidentes de protection de la vie privée, les données nominatives des membres du bureau de l'association[117].

[117] Dans la perspective d'une analyse prosopographique des acteurs associatifs prolongeant l'actuel projet de recherche, ces données nominatives feront l'objet d'une analyse qui permettra de suivre finement les trajectoires militantes de ces acteurs. Rappelons que, d'après l'une des hypothèses de recherche, l'institutionnalisation du phénomène mémoriel aurait provoqué un « effet d'intéressement » auprès d'acteurs associatifs militant auparavant sur d'autres secteurs de revendication ou d'autres domaines d'activité.

2.2 Réduction des données

Nous allons désormais aborder la constitution proprement dite du corpus, après en avoir décrit succinctement la source.
Cette constitution correspond à une véritable construction, dans la mesure où les documents finalement regroupés et traités ont font l'objet de manipulations et de choix multiples. Que désignons-nous par « corpus des déclarations d'associations au JO » ? Nous aimerions insister ici sur la dimension de « fabrication » des données, qui ne sont précisément jamais « données ». Nous avons déjà entr'aperçu certains aspects de cette fabrique (processus auto-déclaratif, sélection des seules associations au sein du JOAFE[118], soustraction des données nominatives dans les fichiers diffusés par la DILA, hétérogénéité de la structuration des fichiers, évolution de la codification des secteurs d'activité attribués aux associations, etc.[119]).
Cette partie sera l'occasion, non plus d'analyser les conditions de collecte et de mise en forme des données du côté des producteurs et diffuseurs, mais de voir en quoi cette notion de « données non données » existe également du côté des utilisateurs, lors de la constitution du corpus.
Par corpus, il faut entendre une collection agencée de fichiers, un ensemble construit en vue d'une certaine finalité. Notre corpus est composé de données combinées, extraites d'un ensemble plus vaste, lui-même distingué d'autres sources. Les résultats issus du traitement de ce corpus ne sont intelligibles qu'à la condition d'être précisément rattachés à leurs conditions de production, et donc aux choix que nous avons effectués, tout au long du processus de sélection, sur les données collectées.
Le projet, tel qu'initialement conçu, consistait (via le traitement lexicométrique des déclarations d'associations du JO, de 1947 à nos jours, sur l'ensemble du territoire national) à apporter des éléments d'information à Sarah Gensburger, quant à la validité de ses hypothèses concernant l'élaboration de la mémoire en question sociale.
En raison de la durée courte du stage (3 mois) mais également de la complexité à traiter des données textuelles, une sélection des données parmi l'intégralité de ces fichiers a dû être opérée, répondant à plusieurs critères et contraintes (certaines externes[120], d'autres propres au projet[121]), visant à définir de façon pertinente le corpus.

Notre méthode de réduction et de construction des données correspond au schéma suivant :



[118] Rappelons que le JOAFE est constitué des déclarations des associations, des associations syndicales de propriétaires, des fondations d'entreprise et des fonds de dotation.
[119] Pour de plus amples développements sur la fabrique des données, notamment leurs conditions de production, qui mêlent dimensions organisationnelles, politiques et techniques, voir DENIS J. et S. GOËTA, « La fabrique des données brutes. Le travail en coulisses de l'open data », *op. cit.*
[120] Comme la disponibilité des fichiers et leur « lisibilité » par les outils d'analyse textuelle.
[121] Ainsi de la pertinence scientifique du choix des mots-clés.

Cette réduction des données s'est faite en plusieurs étapes, de façon itérative. Ces choix ont toujours été fondés sur des discussions que nous avons eues régulièrement avec Sarah Gensburger, afin de s'assurer de leur adéquation aux problématiques de la recherche.

2.2.1 Premier facteur de réduction des données : Accessibilité et qualité des fichiers

Malgré leur intérêt réel, les déclarations publiées au JO représentent un ensemble hétéroclite de fichiers, dont le format natif (papier ou numérique), la source (JO « Lois et décrets » ou JOAFE) et la structuration interne des données varient selon la période considérée. Cette diversité implique des manipulations multiples.
A partir de 1997, les fichiers transmis par la DILA se présentent sous la forme de fichiers XML[122], permettant une transformation aisée en CSV[123] puis en format brut (.txt), manipulable par les outils de textométrie utilisés.
Avant cette date, les fichiers du JO ne sont accessibles que sous la forme de scans océrisés[124] (ou de microfiches archivées à la BnF pour la période 1985-1996), générant de nombreuses erreurs de retranscription. L'importance de ces problèmes et erreurs (structuration des données implicite, graphies non ou mal reconnues[125]) est telle qu'il a été décidé de ne pas utiliser les données publiées avant 1997 dans le cadre de ce stage.
Mathilde de saint-Léger, qui a procédé à une analyse des fichiers scannés (le nombre de pages, composées de plusieurs déclarations, est de 30 746 de 1960 à 1984), a estimé que le coût pour traiter et vérifier les données sur la seule période 1960-1984 s'élève à un montant de 10 mois / homme. La question de l'intérêt et de la rentabilité (en termes de ressources humaines, économiques et matérielles investies au regard de la finalité du projet) du traitement des données ne doit jamais être oubliée ni minimisée dans une démarche d'analyse de données textuelles.

Dans l'optique de futurs travaux s'inscrivant dans le projet de recherche, une sélection de fichiers scannés a tout de même fait l'objet de prétraitements[126], sur la période 1963-1984.

---

[122] « L'*Extensible Markup Language* (XML, « langage à balise extensible » en français) est un langage informatique de balisage générique qui permet de décrire des données, de façon hiérarchique (héritage des propriétés). Cette syntaxe est dite « extensible » car elle permet de définir différents espaces de noms, c'est-à-dire des langages avec chacun leur vocabulaire et leur grammaire. L'objectif initial est de faciliter l'échange automatisé de contenus complexes entre systèmes d'informations hétérogènes (interopérabilité). »,
Page Wikipédia : https://fr.wikipedia.org/wiki/Extensible_Markup_Language
« Langage de balisage extensible basé sur la séparation nette entre le contenu, le style et la structure des documents. XML est un métalangage, ensemble de règles permettant de définir d'autres langages balisés spécialisés. »,
http://www.adbs.fr/xml-extensible-markup-language--19063.htm?RH=OUTILS_VOC

[123] « *Comma-separated values* (CSV) est un format informatique ouvert représentant des données tabulaires sous forme de valeurs séparées par des virgules. », https://fr.wikipedia.org/wiki/Comma-separated_values

[124] « OCR (*Optical Character Recognition* ou, en français, Reconnaissance optique de caractères) : Technique qui repose sur les méthodes appliquées en reconnaissance de formes et qui, à partir d'un procédé optique, permet à un système informatique de lire et de stocker de façon automatique du texte dactylographié, imprimé ou manuscrit sans qu'on ait à retaper ce dernier. La marge d'erreur des systèmes de reconnaissance optique de caractères varie considérablement selon leur degré de perfectionnement. », http://www.granddictionnaire.com/ficheOqlf.aspx?Id_Fiche=2071511

[125] Par exemple, en 1960, le « but » (ou objet, qui est un terme essentiel à la structuration des données textuelles) de l'association est retranscrit par le logiciel de reconnaissance de caractères comme « thn ». Autres exemples, en 1976, le terme « aspirations » est retranscrit en « aspi- • rations » et commémoratives en « comme.-morative,s » Ces erreurs de graphie, outre les nombreux problèmes qu'elles soulèvent pour être résolues, mettent en échec les possibilités de traitement fiable par un outil d'analyse textuelle.

[126] Ces prétraitements consistent en le développement de scripts en Python, effectués par Mathilde de Saint-Léger (structuration des données) et par Elias Benaïssa (nettoyage des données). Elias Benaïssa

Etant donné la diversité des formats de fichiers disponibles – et du coût de traitement afférent –, la couverture du projet scientifique a été adaptée selon un plan de sélection. Celui-ci a été établi selon deux paramètres (période concernée et filtrage par mots-clés[127]).

Voici un tableau synthétique de ce **plan de sélection des fichiers** :

Années et format de fichiers	**Mode de sélection des fichiers**
1940-1959 Fichiers scannés océrisés	Retranscription manuelle des seules déclarations pertinentes (contenant les 6 mots-clés mémoriels)
1960-1969 Fichiers scannés océrisés	Sélection d'1 année toutes les 3 années Fichiers filtrés par 6 mots-clés
1970-1984 Fichiers scannés océrisés	Toutes les années sélectionnées Fichiers filtrés par 6 mots-clés
1985-1996 (Corpus BnF) Microfiches	Toutes les années sélectionnées Fichiers filtrés par 5 mots-clés Couverture géographique (fichiers actuellement traités) : Paris, Isère Extension de la couverture géographique (projet futur) : Bouches-du-Rhône, Cantal, Nantes, Rhône, Hauts-de-Seine, Seine-Saint-Denis, Martinique
1997-2014 Fichiers XML	Toutes les années sélectionnées, Fichiers filtrés par 6 mots-clés

L'accessibilité problématique aux fichiers explique que le choix des fichiers à traiter durant le stage se soit porté sur la période 1997-2014, période pour laquelle le format XML est plus aisément manipulable par des technologies numériques. Mais précisons d'emblée que ce plan de sélection est présenté tel qu'initialement prévu… et que la partie concernant notre stage (1997-2014) a dû également être révisée…

2.2.2 Deuxième facteur de réduction des données : Quantité de données à traiter

Outre l'aspect qualitatif, la quantité de données constitue un autre facteur important à prendre en considération lors de traitement de données. Nous voudrions ici revenir sur une dimension qui nous paraît cruciale dans tout projet d'humanités numériques mais qui semble trop souvent minorée… Celle de la masse de données à gérer – notamment lorsque ces données sont de nature textuelle – et du temps de traitement qu'une telle masse implique.
Nous avions 17 années[128] de fichiers de déclarations (ce qui représente près de 2 millions de déclarations) à notre disposition (1997-2014) ; ce qui, apparemment, ne représente pas (ou, plutôt, n'aurait pas dû représenter), à l'heure et à l'ère du *Big Data*, un défi insurmontable. La couverture temporelle paraissait, en outre, suffisamment importante pour que nous puissions déceler des tendances caractérisant l'évolution du lexique des associations. Nous aurions obtenu *in fine* des résultats qui, bien que n'embrassant pas une période aussi large que prévue, n'auraient pas manqué de profondeur historique. Des phénomènes à

---

a participé au projet collectif « Mémoire », dans le cadre de son stage de fin d'année de Master 1 « Traitement de la langue et Informatique » (Université de Paris-Sorbonne).

[127] Nous développons dans la partie « 4e facteur de réduction des données » le mode de sélection par mots-clés.

[128] Et non 18 années, car le fichier de l'année 2007 n'a pu être intégré dans cet ensemble, en raison de données manquantes sur le site de la DILA.

interpréter, vérifiant ou non les hypothèses de départ, auraient dû émerger « rapidement » de cette masse de données « informes ».

Evaluation quantitative du corpus
Afin de mesurer toute l'étendue du corpus potentiel, nous avons dû procéder à une première évaluation quantitative des fichiers, entre 1963 et 2014, avec une attention particulière portée à la période 1997-2014 (voir en annexe, p. 227, les chiffres détaillés). En voici les résultats sous forme de graphique (nombre total de déclarations par an, sur les deux périodes considérées (sans sélection par mots-clés) :

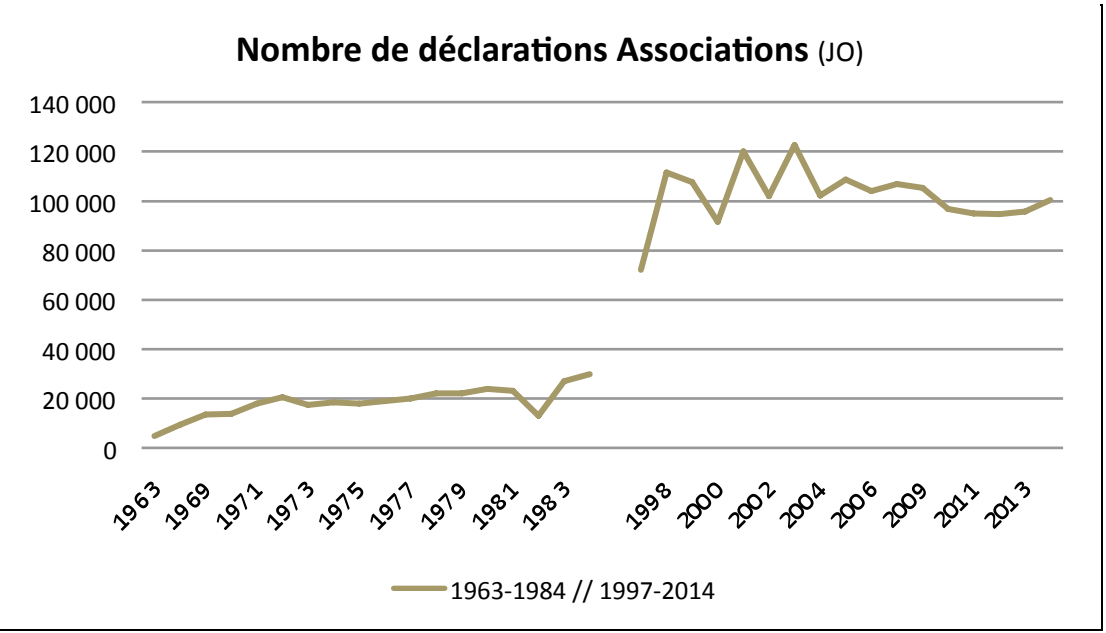


Figure 14 – Evolution temporelle du nombre de déclarations

Cette approche quantitative des fichiers à notre disposition a été nécessaire, non seulement pour la première estimation de l'ampleur du corpus qu'elle a permise, mais également pour la justification des choix futurs dans le processus de sélection des donnés.

Problème de l'aplatissement des données : un accroissement quantitatif majeur
Nous ne présentons ici le problème de l'aplatissement des données – qui sera développé, d'un point de vue technique, dans la partie « Prétraitements » – que pour mettre en avant sa répercussion sur le nombre de déclarations.
Le passage d'un format à un autre (BDD > XML > XSL, TXT) peut entraîner la multiplication des enregistrements pour les champs multivalués. Ce « doublonnage » lié à l'aplatissement des données a occasionné, dans le cas de nos fichiers, un accroissement spectaculaire de l'ordre de 70% ! 737 941 déclarations doublons ont ainsi été artificiellement créées, lors du changement de format. Avant la suppression des doublons (ie, des enregistrements identiques ayant été dupliqués en raison du champ multivalué), le nombre de fichiers entre 1997 et 2014 s'élevait à 2 475 180 pour se trouver réduit à 1 737 239 fichiers, une fois le « dédoublonnage » opéré.
Cette multiplication est due, dans nos fichiers, au champ « Thème », qui correspond à une indexation effectuée à postériori par les services préfectoraux et éventuellement complétée par la DILA, affectant un ou plusieurs domaines d'activité à chaque déclaration. Si cette information peut s'avérer utile pour appréhender la logique d'indexation et de catégorisation des associations par les administrations centrales et les services publics, elle se révèle non pertinente dans le cadre actuel de notre projet dans la mesure où ces données ne sont pas d'ordre auto-déclaratif : ne provenant pas des associations elles-mêmes, elles ne sauraient pour l'instant être intégrées à l'étude du vocabulaire des déclarations. Les fichiers traités ont donc été nettoyés des doublons.

Comment traiter une « inflation associative » ?
Malgré la suppression des doublons, l'évaluation quantitative des fichiers montre une très forte hausse du nombre de déclarations d'associations au JO, de 4 800 en 1963 à plus de 100 000 en 2014. En 2014, le nombre de déclarations correspond ainsi à plus de vingt fois celui de 1963. Loin d'être linéaire, l'évolution du nombre de déclarations connaît des pics (notamment en 1998, 2001 et 2003) et des contractions (2000, 2002 et 2004, par exemple). Précisons que ces chiffres comprennent non seulement les créations d'associations, mais également les modifications (de titre, d'objet ou de siège social) ainsi que les dissolutions.

Cette augmentation du nombre de déclarations, que nous pourrions qualifier d'« inflation associative », représente une réelle difficulté pour qui entend appliquer les techniques de la statistique lexicale. Malgré les indéniables et impressionnants progrès des technologies numériques, vouloir manipuler près de deux millions de fichiers (1 737 239 fichiers) de données textuelles aisément, rapidement et de façon contrôlée, en vue d'un résultat fiable, n'est pas possible – du moins, actuellement. La manipulation textométrique d'un corpus textuel représente un vaste ensemble de micro-opérations (changement de formats, nettoyage et normalisation des données, suppression des doublons, validation du lexique, intégration de variables, etc.) proprement chronophages.
En réduisant le nombre de fichiers sur la période concernée (1997-2014) aux seuls fichiers contenant les mots-clés pertinents, nous obtenons un total de 62 633 déclarations, qui ne peuvent toujours pas être traités ni analysés sur la période limitée de notre stage.
C'est la raison pour laquelle il a été décidé de procéder à une sous-sélection au sein des fichiers. Initialement envisagé, le choix de ne retenir qu'une année sur deux s'est encore avéré trop lourd en termes de traitements et manipulations. Finalement, les années 2000 et 2010 ont été retenues pour être traitées par Calliope et Alceste ; sous-corpus auquel nous avons ensuite adjoint l'année 1984 pour des raisons que nous développerons plus tard.

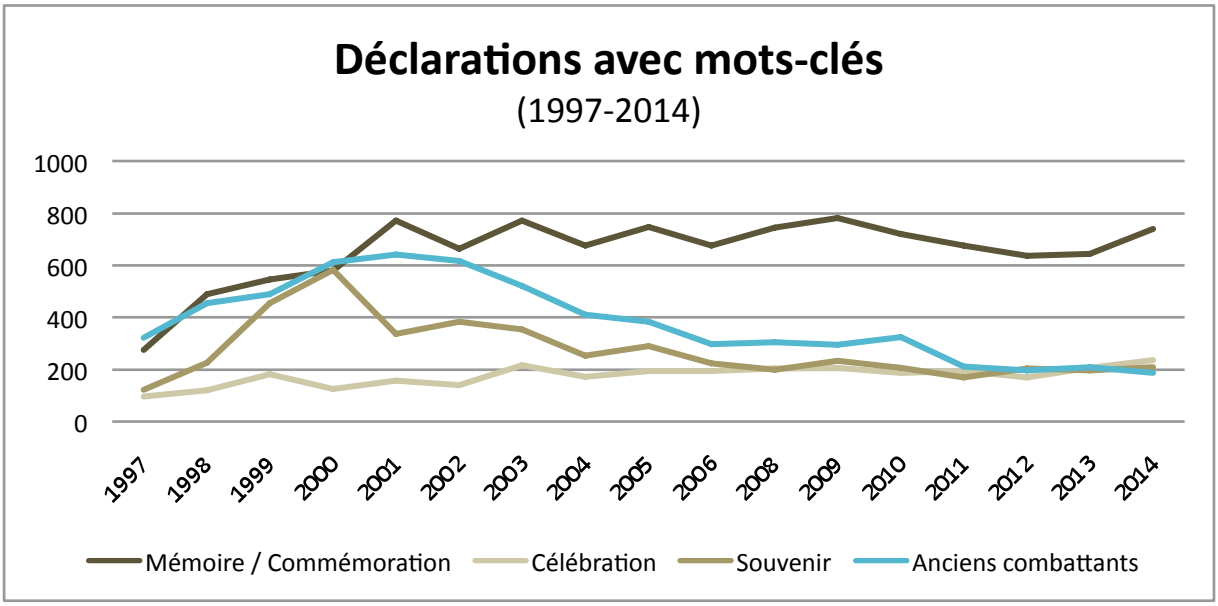
2.2.3 Troisième facteur de réduction des données : Sélection des frontières temporelles

La sélection d'années particulières pour composer le corpus s'est donc imposée en raison de la brièveté du stage effectué (3 mois). Il était évident – et cela le demeure encore – que cette sélection d'années « pertinentes » avait un but exploratoire et qu'elle ne saurait se substituer à une analyse plus approfondie, comprenant davantage de points de référence, sur une période historique élargie, et dont les résultats pourraient confirmer, infirmer ou nuancer ceux auxquels nous sommes parvenus dans le cadre de cette étude. Elle constitue néanmoins une étape liminaire s'inscrivant dans un projet plus large, dont les résultats permettront de mieux cerner les phénomènes à prendre en compte et à analyser dans les étapes ultérieures.

De la difficulté à choisir les fichiers représentatifs
Dès lors que l'exhaustivité n'est plus de mise et que la représentativité de l'échantillon choisi s'impose, émergent de nombreuses questions. Ainsi, quelle méthode de sélection adopter ? Sur quels critères se fonder ? En quoi ces années sont-elles « représentatives » ? Doivent-elles même l'être ? A supposer qu'une année soit retenue en fonction de son lien à tel événement de la politique publique, ne pourrait-on pas nous reprocher de n'avoir conservé et analysé que ce qui pouvait corroborer l'hypothèse de recherche initiale ? Devions-nous laisser le hasard jouer (ou plus justement, le choix non motivé) ? Par ailleurs, une année sélectionnée pour sa pertinence à priori pouvait se révéler, à l'issue du traitement et de l'analyse, pertinente d'un autre point de vue, pour d'autres raisons, en ce sens que les résultats obtenus auraient permis de modifier les hypothèses posées initialement.
Voici quelques-unes des questions qui montrent bien, nous l'espérons, les difficultés qui incombent à quiconque se frotte à la constitution de corpus. Loin d'être une simple formalité (collectionner un ensemble d'éléments), il s'agit ici de réaffirmer le rôle que jouent les choix effectués. La démarche adoptée doit sans cesse questionner les éventuels biais d'interprétation et – dans la mesure du possible – tenter d'y remédier.

> Sélection de l'année 2000 :

Cette sélection de l'année 2000 se fonde sur une évaluation quantitative des fichiers des années 1997-2014. Plusieurs raisons à cela.
Tout d'abord, cette année représente le point (cf. schéma ci-dessous) où les courbes des déclarations contenant les mots-clés « Anciens combattants » et « Mémoire-Commémoration[129] » se croisent pour s'inverser.
De façon générale, le début des années 2000 (2000-2003) représente le point d'inversion de ces deux modalités du registre mémoriel. Avant cette période, l'emploi de ces deux mots-clés semble suivre une évolution parallèle, pour se séparer nettement à partir de 2000-2004. Etudier le vocabulaire des déclarations de cette année, via les techniques de la statistique lexicale, semble donc pertinent pour l'analyse.
Ce phénomène correspondrait, de plus, à l'hypothèse de recherche de Sarah Gensburger : les associations, notamment d'anciens combattants – suivant en cela la démarche du ministère du même nom –, puisent dans le lexique de la mémoire pour asseoir leur pérennité. Nous renvoyons ici à la présentation, en première partie, du projet de recherche et du rôle du ministère des Anciens combattants dans l'institutionnalisation du phénomène mémoriel en France. Il s'agit de voir si se joue un effet de dissémination des termes institutionnels au sein des associations d'anciens combattants. Il faudra bien sûr s'assurer que ce changement de lexique concerne les associations d'anciens combattants ou s'il ne vise pas une réduction du nombre de ces associations, au profit d'autres types d'associations, qui emprunteraient leur vocabulaire au registre mémoriel.
Il faudrait également, pour mieux le saisir, replacer ce premier constat d'une substitution d'un lexique mémoriel à celui des anciens combattants dans son contexte institutionnel et politique, et notamment celui de la création de nouvelles journées nationales de commémoration et de publication de rapports officiels à caractère mémoriel[130].



> Sélection de l'année 2010

[129] Le terme de filtrage qui a été utilisé est « m ?mo », qui a l'avantage d'englober à la fois les termes « Mémoire » et « Commémorations », ainsi que leurs dérivés. Ce choix, qui réduit considérablement le temps alloué aux opérations de filtrage, représente néanmoins l'inconvénient de ne pas pouvoir distinguer ce qui relève du registre mémoriel et du registre commémoratif.
[130] Ainsi, ce ne sont pas moins de 11 journées nationales de commémoration qui ont été instaurées entre 1983 et 2014 et cinq rapports officiels publiés sur les politiques publiques et la mémoire en 2008 – alors que seulement 3 journées nationales de commémoration avaient été créées entre 1914 et 1982. Voir la liste de ces rapports et journées nationales en annexe (p. 221).

Afin de pouvoir comparer l'année 2000 à une autre année – et ainsi analyser la dynamique temporelle des distributions lexicales –, l'année 2010 a également été retenue, cette fois-ci de façon « arbitraire », c'est-à-dire sans fondement quantitatif. Afin de ne pas surdéterminer les résultats en ne retenant que les données à priori « pertinentes » pour l'analyse, nous avons en effet évité de sélectionner une année en raison, par exemple, d'un événement particulier (promulgation d'une loi fixant une journée nationale de commémoration par exemple, ou publication d'un rapport institutionnel). Cela aurait peut-introduit un « biais de confirmation » consistant à ne retenir que les données pouvant vérifier les hypothèses élaborées. C'est pourquoi la deuxième année choisie l'a été hors de toute information extérieure, parmi l'ensemble des années à notre disposition, entre 1997 et 2014. L'écart temporel sélectionné (10 ans) semble, en outre, suffisamment conséquent pour pouvoir appréhender l'évolution, complexe, des usages lexicaux.

> Sélection de l'année 1984

Afin d'élargir le champ d'investigation des données et malgré notre choix initial de ne pas inclure de données issues des fichiers océrisés, une troisième année, antérieure, a été ajoutée à la sélection déjà opérée. L'année 1984[131] a été retenue par défaut – et non 1990, tel que l'on aurait pu s'y attendre, si nous avions reproduire, au sein des fichiers des années précédentes, un intervalle équivalent de dix années. L'année 1990 appartient en effet au lot de fichiers couvrant la période 1985-1996 archivés à la BnF, qui ne sont disponibles que sous forme de microfiches, et donc non traitables dans le cadre de ce stage.

2.2.4 Quatrième facteur de réduction des données : Sélection en fonction des mots-clés

Cette sélection s'avère particulièrement essentielle pour le projet auquel nous collaborons. En effet, les étapes de sélection précédentes, importantes en elles-mêmes, étaient en fait imposées par des contraintes externes au projet de recherche : accessibilité problématique aux données et traitement des données fonction de la durée limitée du stage. La sélection des fichiers par mots-clés répond, quant à elle, à un objectif scientifique : étudier les déclarations d'associations en lien à la question mémorielle. C'est pourquoi la sélection des annonces, parmi l'intégralité des fichiers disponibles, en fonction de leur pertinence pour la problématique visée est capitale.
Nous avons donc opéré une autre sélection en indexant les fichiers au moyen de mots-clés – mots-clés fondés sur les hypothèses élaborées par Sarah Gensburger, exposées en première partie de ce mémoire.
La réduction des déclarations par mots-clés se base également sur la description quantitative des fichiers que nous avons faite. Celle-ci montre en effet une proportion prédominante de déclarations ne relevant pas explicitement de la thématique mémorielle (ie, ne contenant aucun des mots-clés mémoriels[132]). Ainsi, sur la période 1997-2014, les déclarations à « tonalité mémorielle » ne dépassent jamais les 4,3% du total des déclarations. Etudier l'ensemble des déclarations n'aurait donc aucun intérêt pour l'analyse car cela reviendrait à « noyer » les annonces pertinentes parmi une masse de déclarations « hors sujet ».
L'emploi de ces mots-clés, qui représentent une partie de l'univers sémantique de la mémoire, est destiné à « capturer[133] » les déclarations ayant trait, d'une manière ou d'une

[131] Afin d'être opérationnel et manipulable par les outils de lexicométrie, le fichier de l'année 1984 a fait l'objet de prétraitements approfondis de la part d'Elias Benaïssa.
[132] Nous voudrions insister ici sur le fait que la « capture » du phénomène mémoriel au moyen de termes l'exprimant, le désignant n'est pas une question aisée. Il est évident que les mots-clés sélectionnés ne prétendent pas épuiser le vocabulaire mémoriel et que l'analyse lexicométrique des fichiers les contenant vise seulement à permettre de fournir une aide au questionnement scientifique.
[133] Là encore, émerge une difficulté d'ordre méthodologique. Comme nous l'avons souligné plus haut, comment appréhender, d'un point de vue linguistique, la question mémorielle ? Comment capter le champ sémantique de la mémoire sans pouvoir l'épuiser linguistiquement ? Pourquoi le choix de ces mots-clés ? Rappelons que la démarche et l'analyse que nous proposons ici ne valent qu'à titre

autre, à la thématique mémorielle. Une analyse de la distribution de leurs mots et des relations qu'ils entretiennent doit permettre de mieux comprendre comment se décline la question mémorielle, quelles sont les thématiques stables, en développement ou en régression, quels sont les types de discours qui s'y jouent, quelle est sa distribution géographique, etc.

Présentation des mots-clés

Initialement, six mots-clés ont été sélectionnés par Sarah Gensburger pour filtrer[134] les déclarations pertinentes du JO. Le choix de cinq premiers mots-clés (« mémoire », « anciens combattants », « commémoration », « souvenir », « célébration ») repose sur des projets de recherches antérieures qui ont mis en avant leur avènement et leur emploi régulier depuis les années 80, notamment au sein des ministères des Anciens combattants et de la Culture.

Rappelons rapidement le contexte institutionnel et politique qui explique le choix de ces mots-clés :

- En 1920, la mise en œuvre des deux premières journées nationales de **commémoration** créées est confiée au ministère des Pensions, Primes et Allocations de guerre (politique du "**souvenir**" des **anciens combattants** morts pour la Patrie).
- En 1945, un ministère des Anciens combattants et Victimes de guerre (MACVG) est instauré, qui supervise notamment la création et l'entretien des cimetières de guerre et de monuments commémoratifs.
- En 1982, le ministère des Anciens combattants crée la Mission permanente aux commémorations et à l'information historique.
- En 1992, La Mission devient la Délégation à la **mémoire** et à l'information historique
- En 1997, le MACVG devient un simple secrétariat d'Etat (SEAC) rattaché au ministère de la Défense. Les "actions de mémoire" sont alors affichées comme une de ses raisons d'être ;
un Haut Conseil à la Mémoire Combattante est créé pour piloter ce qui est désormais systématiquement qualifiée de "politique de la mémoire".
- En 1999, la Délégation à la mémoire et à l'information historique change de statut et devient une direction d'administration, la Direction de la mémoire, du patrimoine et des archives (DMPA).
- En parallèle, le ministère de la Culture s'empare également du phénomène mémoriel : en 1974, création d'un comité aux **Célébrations** nationales, dont l'action se transformera au fil des années 90 pour s'institutionnaliser autour des termes de "mémoire" et de "commémoration).

Un dernier mot-clé « Patrimoine » sera ajouté aux cinq autres mots-clés, afin d'élargir le spectre mémoriel.

Ces six mots-clés sont donc composés de : « Anciens combattants », « Mémoire », « Célébration », Commémoration », « Souvenir » et « Patrimoine ».

Le filtrage par mots-clés prend également en compte leurs dérivés. Ainsi « Mémoire » rassemble les termes « mémoires », « mémoriel » et mémorielle » ; « Commémoration » indexe les déclarations contenant « commémoratif », « commémorative », « commémorer » ; « Célébration » pour « célébrations », « célébrer », « célébrons », et « Patrimoine » pour « patrimoines », « patrimonial », etc.

A noter une difficulté qui sera développée dans la partie « Prétraitements des données » : celle concernant le traitement du fichier de l'année 1984. Son océrisation a en effet généré de nombreuses erreurs de graphies, qui provoquent une mauvaise reconnaissance des termes lors du filtrage. C'est pourquoi M. de Saint-Léger a dû, à l'aide de Calliope, annoter les différentes graphies rencontrées dans le fichier.

Nous donnons ci-dessous, à titre d'exemple, les variantes rencontrées pour « Anciens combattants » :

---

temporaire, exploratoire, et dont les résultats produits permettront d'affiner ensuite la méthode et les choix effectués.

[134] Ce filtrage a été obtenu via la fonction d'indexation de Calliope qui permet, de façon relativement aisée et rapide, de filtrer les déclarations pertinentes.

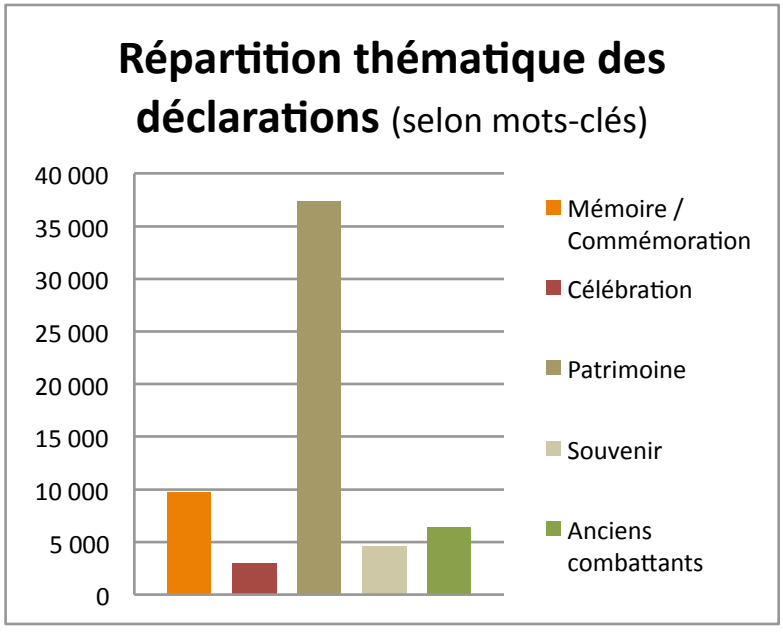
ANCIENS-COMBATTANTSABATTANT ANCIEHS ANCIEN COMBALTANTS ANCIEN COMBAT ANCIEN COMBATIANTS ANCIEN COMBATTANT ANCIEN COMBATTANTI ANCIENS-COMBATTANTS-D ASSIENS COMBATTANT CAMBATTANTS CEMBATTANTS COINBATTANTS COMBATTANB COMBATTANIS COMBATTANS COMBATTANST COMBATTANT COMBATTANT% COMBATTANTE COMBATTANTI COMBATTANTS- -COMBATTANTS COMBATTANTS-DU COMBATTANTS-UNION COMBATTINTS COMBATTONTS COMUATTANTS CONKBATTANTS EMBATTANTS NCIENS-COMBATTANTS

Le filtrage des déclarations par mots-clés a représenté une forte réduction du nombre de documents à traiter, comme le montre le tableau ci-dessous[135] :

	1984	**2000**	**2010**
Nombre total de déclarations	29 900	91 557	96 803
Nombre de déclarations pertinentes (avec mots-clés)	1 397	2 747	3 368

Suppression des mots-clés « Patrimoine » et « Célébration »
Un autre processus de réduction des données a pu être effectué, grâce à une analyse quantitative des fichiers pertinents (c'est-à-dire contenant les six mots-clés). Un décompte des fichiers filtrés révèle en effet une inadéquation des mots-clés « Patrimoine » et « Célébration » à l'objectif visé.

Deux raisons expliquent le rejet des déclarations contenant ces deux mots-clés dans le cadre de notre étude :
- **Poids prépondérant du mot-clé « Patrimoine »**
Les déclarations contenant le mot-clé « Patrimoine », toutes années confondues (1997-2014), sont largement majoritaires et écrasent, du fait de leur poids, les autres déclarations. Ainsi, 60 % des déclarations pertinentes contiennent le mot-clé « Patrimoine » et ses dérivés, représentant à lui seul près des deux-tiers de l'ensemble des déclarations, comme le montre le graphique ci-dessous :



Le vocabulaire patrimonial, surreprésenté au sein du

[135] Voir en annexe la distribution quantitative des déclarations pertinentes de 1997 à 2014 (p. 228).

lexique global, risquait d'introduire une sorte de biais dans l'étude : les nuances lexicales moins fréquentes pouvaient devenir invisibles du fait de sa prédominance. C'est pourquoi il a été décidé, dans un premier temps, de procéder à trois types de traitements distincts : un traitement global du vocabulaire suivi du traitement du seul lexique patrimonial et de celui du lexique hors patrimoine, de manière à voir les changements opérés. Nous aurions pu, en effet, traiter l'ensemble des déclarations pertinentes, pour ensuite distinguer le corpus en sous-corpus, afin d'analyser finement les caractéristiques lexicales de chacun et de procéder à leur comparaison. Outre des questions de temps, l'abandon de cette solution est surtout motivée par un manque de pertinence de ces deux mots-clés relativement à la question mémorielle.

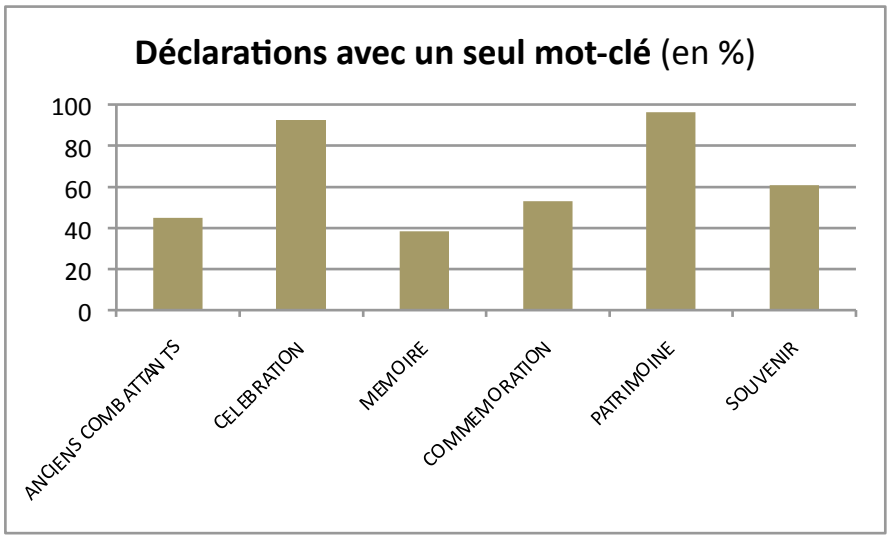
- « Patrimoine » et « Célébration » : des thématiques disjointes

La (sur)présence du mot-clé « Patrimoine » relativement aux autres mots-clés n'aurait pas, à elle seule, légitimé son exclusion si un autre aspect n'était venu s'y ajouter.

Nous avons en effet étudié la combinaison des mots-clés entre eux pour déterminer quelles étaient les possibles zones de chevauchement, d'entrelacement thématique ou, au contraire, les thèmes disjoints, n'entretenant que peu de relations aux autres thèmes. Il s'avère que les mots-clés « Patrimoine » et « Célébration » apparaissent, notamment en 2000, de manière quasi exclusive dans les déclarations (au sein des rubriques « Titre » et « Objet »), ne cooccurrant que très faiblement avec les autres mots-clés.

Cette analyse des cooccurrences des mots-clés a ainsi permis de montrer que, dans la perspective de la constitution du corpus final, sans contextualisation a minima des termes, leur sens d'emploi est impossible à appréhender. La sélection des termes « Patrimoine » et « Célébration » paraissaient en effet pertinents pour l'analyse en début de processus. L'analyse des cooccurrences permet d'affiner la sélection des données, d'en opérer une réduction – en cela, elle se distingue d'un processus par échantillonnage aléatoire à partir du terme "patrimoine, qui viserait à vérifier ce à quoi renvoie ce terme.

A titre d'exemple, voici quelques résultats issus de l'année 2000 :



La proportion d'apparition des mots-clés seuls (ie, sans combinaison avec d'autres mots-clés) au sein des déclarations est quasi totale pour les mots-clés « Patrimoine » (96,47% des fichiers contenant « Patrimoine » ne contiennent pas d'autres mots-clés) et « Célébration » (92,42%).

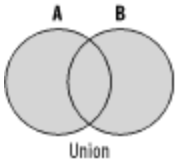
A l'inverse, le mot-clé « Mémoire » est celui qui cooccurre le plus avec les autres mots-clés (seuls 38,37% des fichiers de l'année 2000 contiennent uniquement le mot-clé « Mémoire »).

Situés entre ces deux pôles « mémoriels », l'un constitué de thèmes fortement isolés, l'autre d'un thème largement partagé, nous trouvons les mots-clés « Anciens combattants », majoritairement lié à d'autres thèmes (45% des fichiers contiennent uniquement le mot-clé « Anciens combattants ») ; le mot-clé « Commémoration », qui apparaît seul dans 53% des fichiers et, enfin, le mot-clé « Souvenir », avec 61%.
Sans vouloir préjuger des résultats, il nous semble néanmoins intéressant de noter que les deux mots-clés les plus liés aux autres mots-clés soient « Mémoire » et « Anciens combattants ».

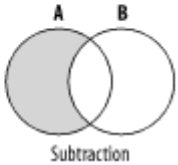
Les thèmes du patrimoine (relevant surtout du domaine culturel et environnemental[136]) et de la célébration (appartenant principalement à la sphère religieuse et cultuelle) constituent des domaines de la mémoire isolés, qui n'entretiennent que très peu de liens avec les autres thématiques de l'objet mémoriel nous intéressant (liens entre les anciens combattants et la mémoire). Une analyse de leur vocabulaire, certes intéressante en soi, n'apporterait pas d'éléments significatifs à la compréhension des thématiques mémorielles qui nous préoccupent, tant leur univers thématique est singulier.
Les déclarations indexées par ces mots-clés n'ont donc pas été retenues dans notre étude.

Une analyse nécessairement conjointe de l'univers thématique des mots-clés
De la même manière que nous souhaitions comparer les liens lexicaux entre les déclarations patrimoniales et celles mémorielles, au sens large, nous aurions voulu analyser, conjointement et séparément, les déclarations en fonction de leurs mots-clés. Le processus aurait été le suivant :

- L'analyse de l'ensemble des déclarations aurait permis d'en distinguer les thèmes majeurs et mineurs (A U B).



- L'analyse distincte des déclarations en lien à chaque mot-clé, sans les autres mots-clés (par exemple, celles contenant « mémoire » sans « anciens combattants », etc.) aurait permis de repérer les thématiques propres à chaque registre mémoriel, d'en révéler les nuances lexicales, et notamment celles de la "mémoire". (A \ B)



- La comparaison des deux types de traitements aurait mis en valeur les caractéristiques propres à chaque groupe ainsi que les éventuelles influences intergroupes.
Malheureusement, pour des questions de trop faible quantité de données – notamment pour l'année 1984, dont l'effectif est beaucoup plus faible que celui des autres années –, cette solution n'a pu être retenue. Une certaine quantité de données doit en effet être réunie pour que son traitement statistique soit pertinent.

[136] Ce thème comprend essentiellement la protection du patrimoine (historique, architectural, industriel, culturel, etc.) et celle de l'environnement (protection de la biodiversité notamment).

Mais un moyen de contourner cette difficulté s'esquisse déjà, dans la perspective de futurs travaux : opérer des ponctions de déclarations, tout au long de la période (du moins, de 1963[137] à 2014) qui seraient regroupées en « paquets » d'années, de façon à ce que le seuil minimal pour un traitement statistique valide soit atteint. Ainsi, si le seuil correspond à une centaine de déclarations (seuil qui reste à déterminer) et que, pour atteindre ce seuil en début de période, il faille regrouper les fichiers de trois années (par exemple, 1963, 1964 et 1965), alors les autres groupes de fichiers seraient également composés de trois années. Ces ponctions seraient effectuées à intervalle régulier (tous les 5 ans, donc dans notre exemple, à partir de 1970), de manière à ce que les ponctions demeurent bien des ponctions et ne reviennent pas à analyser chaque année.

2.2.5 Cinquième facteur de réduction des données : Sélection en fonction du type d'annonces au JO

Dernier facteur de réduction des données, la sélection des fichiers en fonction du type d'annonces publiées au JO.
Nous avons présenté dans la première partie, la DTD[138] qui présidé à la structuration des fichiers de la DILA (depuis 1997). Ce document de référence définit les différents types de classes décrivant les données des déclarations ainsi que leur agencement.
L'élément « Type » de cette DTD informe sur les différentes types d'annonces publiées, dont chaque modalité est pourvue d'une valeur spécifique. Ainsi, la modalité « Création d'association » correspond à la valeur « 1 », la « Modification d'association » à la valeur « 2 » et la « Dissolution d'association » à la valeur « 3 ». Chacune de ces modalités comprend des sous-modalités, telles que « 111 » qui caractérise une « Annulation de création d'association ».

Type d'annonce
1 Création d'association
11 Rectificatif de création d'association
111 Annulation de création d'association
2 Modification d'association
22 Rectificatif de modification d'association
222 Annulation de modification d'association
3 Dissolution d'association
33 Rectificatif de dissolution d'association
333 Annulation

Les modifications concernant les sièges sociaux ont été exclues de notre corpus. En effet, cette information, importante dans le cadre d'une analyse prosopographique – dans la mesure où la domiciliation du siège social est souvent liée à la résidence du président de l'association –, n'apporte pas d'élément pertinent à une analyse du lexique employé par les associations. Notre étude s'est donc focalisée sur les rubriques « Titre » et « Objet » (ainsi que leurs corollaires « Nouveau titre » et « Nouvel objet »).
La spécification du type de modifications dont font l'objet les annonces du JO varie selon l'année considérée. En 1984, les divers types de modifications sont explicitement mentionnés au sein d'une rubrique spécifique (par exemple, « L'association X transfère son siège social de… à … » ou encore « L'association Y change son intitulé »), dans la version papier du *Journal officiel*. Pour les années 2000 et 2010, le changement d'adresse doit être déduit par filtrage des autres rubriques. Seul figure, dans les fichiers, un code indiquant qu'une

[137] Les fichiers des années datant des années 1963 à 1983 ont en effet fait l'objet d'une structuration et d'un nettoyage, comme énoncé plus haut, lors du stage d'Elias Benaïssa. Ils sont donc désormais disponibles pour des traitements de statistique textuelle.
[138] Pour une présentation détaillée de la DTD, se reporter à la partie « Description de la source » et à l'annexe correspondante (p. 223).

modification a eu lieu mais non de quel ordre. Afin d'exclure du corpus les changements de siège social, nous n'avons retenu, parmi les déclarations ayant fait l'objet d'une modification, que celles qui ont plus qu'un titre seul. Autrement dit, si les rubriques « Objet », « Nouvel objet » et « Nouveau titre » d'une déclaration sont vides, nous pouvons alors en déduire qu'il s'agit d'un changement de siège social, et éliminer la déclaration correspondante.

Dernière précision, ont également été exclues du corpus à analyser les annonces qui enregistrent les dissolutions d'associations. En effet, le projet avait pour but, dans un premier temps, d'étudier sur le long terme le lexique des associations, nouvellement créées ou celles déjà existantes qui modifiaient leur contenu. La disparition d'associations ne correspondait donc pas à cette finalité. Une évaluation quantitative de ces suppressions sera néanmoins faite, ainsi qu'une analyse du type d'associations, afin de voir si des associations en lien à la thématique mémorielle (notamment, celles d'anciens combattants) sont concernées. Si tel est le cas, une discussion sur leur éventuelle réintroduction au sein des données à traiter aura alors lieu.

2.3 Le corpus final : nombre et type de fichiers traités

Afin de conclure cette partie sur la constitution du corpus, nous présentons dans le tableau ci-dessous un résumé de l'ensemble des différentes opérations de sélection des données que nous avons effectuées, avec leur traduction quantitative.

Tableau récapitulatif du nombre de déclarations à l'issue du processus de sélection :

	1984	**2000**	**2010**
Nombre de déclarations initial	29 900[139]	153 738	123 030
Nombre de fichiers océrisés	2 469	X	X
Nombre de déclarations sans doublons	X	91 557	96 803
Nombre de déclarations contenant au moins l'un des 6 mots-clés	1 397	2 747	3 368
Nombre de déclarations contenant au moins l'un des 4 mots-clés	655	1 394	937
Nombre de déclarations selon types d'annonces (sans dissolution)	649	1 323	855

[139] Ce chiffre est une estimation correspondant au nombre de déclarations obtenues à partir des fichiers identifiés comme contenant un des mots-clés mémoriels.

Nombre de déclarations hors changement de siège social	587	1 107	707

3 Prétraitements des données

Après avoir décrit la méthode qui a guidé la composition de notre corpus, nous allons désormais présenter les différentes étapes de transformations des données nécessaires à leur traitement par des outils d'analyse textuelle. Précisons que toutes les opérations décrites ici n'ont pas été appliquées dans le cadre de ce stage. Elles sont néanmoins abordées en raison de leur utilité en vue de futures analyses et dans la mesure où notre travail s'insère dans un projet collectif plus large, englobant différentes personnes, différentes compétences.

Avant de les détailler, voici une vue synthétique des transformations nécessaires au traitement par les outils d'analyse textuelle :

Prétraitements : nettoyage et formatage des données en vue du traitement					
	Calliope	**Alceste**	**Procédés utilisés / automatisation**	**Finalité**	**Compétences requises**[140]
Mise au format des données					
Transformation BDD / XML > format brut (.txt)	X	X	Editeur de texte en format brut	Traitement par outils ADT	1
Création de lignes et mots étoilés		X	Word OpenRefine[141]	Analyse croisée des déclarations via variables illustratives	1
Structuration des données					
Homogénéisation structure	X	X	Scripts Python[142] Excel	Comparaison entre fichiers	3 / 1
Nettoyage et normalisation des données / Mise en forme					
Correction des graphies défectueuses	x	X	Scripts Python	Traitement par outils ADT	3

[140] Par niveaux de compétences requis, nous entendons les connaissances et savoir-faire nécessaires à la réalisation d'un certain processus. Ces niveaux se déclinent comme suit : 1) Aucune compétence particulière autre que la maîtrise de la suite bureautique ; 2) Compétences aisément accessibles ; 3) Compétences requérant un certain temps pour leur acquisition et maîtrise ; 4) Expertise nécessaire.

[141] OpenRefine est une application libre dédiée au profilage et au nettoyage automatisé des données. Il a été conçu pour traiter de larges jeux de données tabulaires, importables et exportables sous une variété de formats (TSV, CSV, Excel, XML, JSON, RDF). Voir le site : http://openrefine.org.
Pour une présentation détaillée d'OpenRefine et de ses fonctionnalités, voir HOOLAND S. van et R. VERBORGH, *Linked data for librarie*s, *archives and museums: how to clean, link and publish your metadata*, London, Facet Publishing, 2014., et VERBORGH R. et M. DE WILDE, *Using OpenRefine. The essential OpenRefine guide*, Packt Publishing, 2013.

[142] Comme nous l'avons déjà dit, les scripts Python ont été développés par Mathilde de Saint-Léger et Elias Benaïssa, pour nettoyer les fichiers de 1963-1984 qui seront traités ultérieurement, dans le cadre du prolongement du projet de recherche. Le langage Python est un « langage de programmation objet, multi-paradigme et multiplateformes. Il favorise la programmation impérative structurée, fonctionnelle et orientée objet. Il est particulièrement utilisé comme langage de script pour automatiser des tâches simples mais fastidieuses, », Page Wikipédia : https://fr.wikipedia.org/wiki/Python_%28langage%29

Transformations des caractères (majuscule > minuscule, suppression espaces inutiles, traits d'union > tiret bas, etc.)		X	Expressions régulières OpenRefine	Traitement par outils ADT	2
Typage (format date, chiffre, etc.)	X	X	Expressions régulières OpenRefine	Création de variables illustratives	2
Homogénéisation des désignations		X	Fonctionnalité « Clusterisation »OpenRefine	Création de variables illustratives	2

Qu'est-ce que le prétraitement des données ?
Le prétraitement des données consiste à appliquer aux données différentes opérations visant à les transformer en vue de leur utilisation par les outils d'analyse textuelle que sont Calliope et Alceste. Précisons que l'utilisation d'Alceste (comme le montre le tableau ci-dessus) nécessite davantage de prétraitement, et donc de temps que celle Calliope. Ce facteur – non négligeable – est à prendre en considération et doit être estimé sur le plan quantitatif dans tout projet de traitement textométrique de données.
Autre précision, la réalisation de ces opérations de prétraitements fait appel à des compétences variées et de niveaux différents. Autant l'emploi de la fonction « rechercher-remplacer » de Word ou celles de calcul de texte d'Excel est aisée, autant le développement de scripts en langage de programmation (Python ou Java) requiert des compétences certaines[143]. Située entre les deux, l'utilisation d'expressions régulières d'Open Refine demande un temps, relativement petit, de formation. Développer un projet d'analyse de données textuelles, c'est aussi et déjà s'assurer que les compétences et savoir-faire nécessaires peuvent être réunis.

(Semi)-automatisation du processus de prétraitement
Ces prétraitements sont constitués d'un ensemble de micro-opérations qui, outre le temps qu'elles nécessitent, représentent autant de sources potentielles d'erreurs de manipulation. C'est pourquoi chaque étape doit non seulement être justifiée mais également reposer sur une méthode appliquée tout au long du processus, qui nécessite parfois – à l'instar de la constitution du corpus – de revenir en arrière pour choisir une nouvelle orientation.
Afin de limiter le temps dévolu aux prétraitements et les sources d'erreurs manuelles, une automatisation des processus a été recherchée, à chaque fois que possible. Ainsi, une partie des prétraitements a pu être automatisée (notamment, pour la vérification de la qualité des données, leur structuration et leur nettoyage) via des scripts et des expressions régulières. Mais cette automatisation requérant elle-même du temps, des compétences particulières et, parfois, des vérifications manuelles finales, la pertinence pour le projet de traiter tel fichier doit à chaque fois être justifiée.

« Profilage » des données[144]
Avant de pouvoir manipuler des données et de les soumettre à des outils d'analyse textuelle, il convient de s'assurer de la qualité et de la cohérence de ces données. C'est pourquoi la première étape de prétraitement nécessite l'établissement d'un « profilage » des données. Ce profilage consiste à poser un diagnostic général sur la qualité et la cohérence des

[143] Nous tenons à préciser que le développement de ces scripts, qui excèdent nos propres compétences, a été effectué par d'autres personnes. Preuve, s'il en est, de la dimension collective inhérente à tout projet d'humanités numériques et à la nécessaire répartition des compétences et des tâches qu'elles impliquent dans leur déploiement.
[144] La notion de « profilage des données » est définie par Olson (cité dans *Linked Data, op. cit.*, p. 77) comme « the use of analytical techniques to discover the true structure, content and quality of a collection of data ».

données, en vue de leur traitement. Il permet ensuite de décider des actions à prendre afin d'améliorer la qualité des données.
La compilation de données issues de différentes sources ou publiées à différentes périodes engendre, comme c'est le cas avec les fichiers des déclarations du JO, des incohérences au sein des données (variété des formats de dates, des structures de données, des graphies, etc.) qu'il s'agit de repérer avant de pouvoir les modifier.
Les fonctions d'OpenRefine, telles que les facets et les filtres, facilitent cette vérification ainsi que les actions de modification (utilisation d'expressions régulières pour automatiser le processus de conversion).

3.1 Opérations de prétraitements des données

Une fois le périmètre et la nature des incohérences affectant les données clairement définis, il est alors possible de procéder aux modifications proprement dites.
Ces opérations se déclinent en trois catégories, dont les deux premières s'appliquent indifféremment aux deux outils utilisés, Calliope et Alceste – tandis qu'une partie de la troisième ne concerne qu'Alceste :
- Mise au format des fichiers,
- Structuration des données,
- Nettoyage et normalisation des données.

3.1.1 Mise au format des fichiers

Pour tout traitement lexicométrique, il faut veiller à ce que les fichiers soient dans un format compatible à celui des outils (.txt). Les fichiers XML (exportés à partir de BDD) fournis par la DILA doivent donc être transformés en fichiers .txt. L'aplatissement des données hiérarchiquement structurées (XML) en données tabulaires (.txt) nécessite de créer, sous Excel, un mappage XML, permettant d'associer à chaque colonne des fichiers l'entité XML correspondante. Concrètement, il s'agit de conserver une seule et même structure pour l'ensemble des données et de l'associer aux 52 fichiers[145] qui constituent une année de déclaration au JO (un fichier pour chaque semaine de l'année).

« L'effet papillon » de l'aplatissement des données
Le passage d'un modèle de données relationnel ou hiérarchique en données tabulaires peut générer des changements dans l'agencement des données loin d'être anodins. Ainsi, les valeurs d'un champ multivalué d'une base de données sont, une fois aplaties en format tabulaire, transformées en autant de lignes supplémentaires que le champ a de valeurs. Un enregistrement comportant x valeurs pour un même champ sera décliné en x enregistrements au sein du fichier text, les informations caractérisant ces enregistrements étant identiques, de ligne à ligne, à l'exception des valeurs du champ multivalué, différentes selon les lignes.
Comme nous l'avons vu, cet aplatissement est responsable d'un accroissement de plus de 70% des enregistrements. Cette multiplication d'enregistrements concernant essentiellement, dans nos fichiers, le champ « Thème » de la DTD de la DILA (qui définit les différents secteurs d'activité des associations), nous avons choisi, pour des questions de limites temporelles, de supprimer les lignes surnuméraires. Les informations contenues dans le champ « Thème » n'ont donc pas été conservées dans le cadre de notre stage.
Mais dans l'optique de futurs travaux (analyse de la représentation des associations par les administrations centrales), ces informations ont néanmoins fait l'objet d'un traitement particulier. Un programme en Python a ainsi été développé de manière à conserver l'ensemble des valeurs de ce champ sans multiplier le nombre d'enregistrements.

[145] Le nombre de ces fichiers n'est valable que pour la période 2004-2014. Avant cette période, la périodicité de publication du JO étant quotidienne, le nombre de fichiers augmente d'autant.

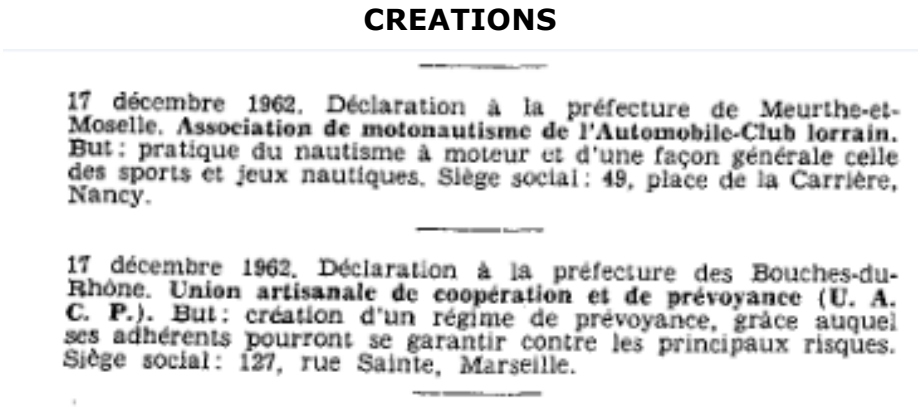
Par ailleurs, les fichiers .txt doivent être encodés en UTF-8[146], en raison notamment du caractère universel et largement compatible de ce type de codage. Les caractères, notamment accentués ou spéciaux, seront ainsi plus facilement reconnus et interprétés par les différents outils numériques utilisés, présents ou à venir.

3.1.2 Structuration des données

Disposer de données structurées, qui plus est de façon homogène, constitue une étape incontournable dans la manipulation de données. Outre la possibilité de pouvoir traiter et analyser ces données, une structuration homogène rend possible la comparaison de plusieurs fichiers entre eux, sur différentes périodes.
Les problèmes que soulève la structuration des données sont de deux ordres : mettre en place une structuration lorsque celle-ci est absente ou non manipulable par les outils informatiques ; pallier les incohérences qu'une structuration perdurant dans le temps ne manque pas de générer. Ces deux aspects peuvent se combiner[147] mais pour des raisons méthodologiques, nous les distinguerons ici.

Mise en place d'une structuration des données
Selon les dates de création des fichiers, et selon leur format natif (papier ou électronique), l'opération de structuration et d'homogénéisation est plus ou moins complexe.
Pour les fichiers antérieurs à 1985, les fichiers sont constitués de scans des JO papier (1947-1984) ou de microfiches (1985-1996). Les fichiers scannés océrisés se présentent de façon non structurée – simples suite de données textuelles –, et ont donc dû faire l'objet d'une structuration, sous la forme de développement de scripts en Python. Ces scripts ont ainsi permis d'automatiser, pour partie, le processus.

Voici des exemples de fichiers scannés, datant de l'année 1963.



CREATIONS

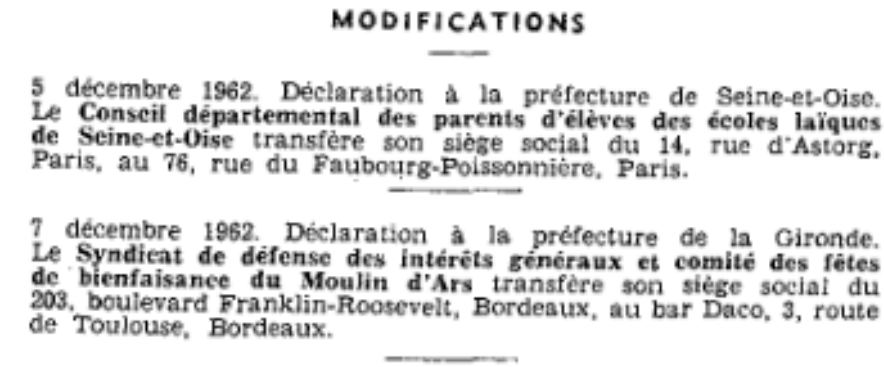
17 décembre 1962. Déclaration à la préfecture de Meurthe-et-Moselle. **Association de motonautisme de l'Automobile-Club lorrain.** But : pratique du nautisme à moteur et d'une façon générale celle des sports et jeux nautiques. Siège social : 49, place de la Carrière, Nancy.

17 décembre 1962. Déclaration à la préfecture des Bouches-du-Rhône. **Union artisanale de coopération et de prévoyance (U. A. C. P.).** But : création d'un régime de prévoyance, grâce auquel ses adhérents pourront se garantir contre les principaux risques. Siège social : 127, rue Sainte, Marseille.

Figure 18 – Exemple de créations

[146] Page Wikipédia « UTF-8 » : « L'UTF-8 (Universal Character Set Transformation Format - 8 bits) est un codage de caractères informatiques conçu pour coder l'ensemble des caractères du « répertoire universel de caractères codés ». De par sa nature, UTF-8 est d'un usage de plus en plus courant sur Internet, et dans les systèmes devant échanger de l'information. Il s'agit du codage le plus utilisé dans les systèmes GNU, Linux et compatibles pour gérer le plus simplement possible des textes et leurs traductions dans tous les systèmes d'écritures et tous les alphabets du monde. » , https://fr.wikipedia.org/wiki/UTF-8

[147] Ce qui est d'ailleurs le cas du projet collectif dans lequel nous nous insérons : il a fallu en effet structurer les fichiers océrisés et veiller à ce que cette structure soit cohérente et homogène quelle que soit l'année concernée.

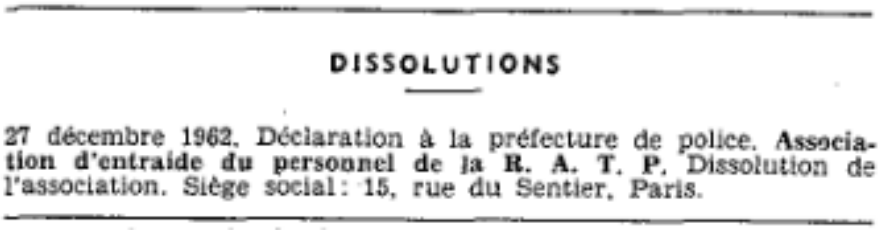


MODIFICATIONS

5 décembre 1962. Déclaration à la préfecture de Seine-et-Oise. **Le Conseil départemental des parents d'élèves des écoles laïques de Seine-et-Oise** transfère son siège social du 14, rue d'Astorg, Paris, au 76, rue du Faubourg-Poissonnière, Paris.

7 décembre 1962. Déclaration à la préfecture de la Gironde. **Le Syndicat de défense des intérêts généraux et comité des fêtes de bienfaisance du Moulin d'Ars** transfère son siège social du 203, boulevard Franklin-Roosevelt, Bordeaux, au bar Daco, 3, route de Toulouse, Bordeaux.

Figure 19 – Exemple de modifications



DISSOLUTIONS

27 décembre 1962. Déclaration à la préfecture de police. **Association d'entraide du personnel de la R. A. T. P.** Dissolution de l'association. Siège social : 15, rue du Sentier, Paris.

Figure 20 – Exemple de dissolutions

Une structure implicite à expliciter

La structuration, implicite, du contenu du JO apparaît au moyen d'éléments langagiers, de conventions graphiques et typographiques, compréhensibles par les lecteurs humains mais non traitables tels quels par des machines. D'un point de vue « informatique », ces éléments océrisés ne sont en effet que des signes, des caractères, sans signification, et auxquels il faut, en quelque sorte, associer une syntaxe et une sémantique, au moyen de scripts, visant à ordonner et « in-former » cette masse de caractères.

Une mise en page particulière distingue les déclarations entre elles (filet court) ainsi que le type d'annonces (filet long, titre en capitales et en gras pour les créations, modifications et dissolutions).

Au sein de chaque déclaration, la structure implicite se révèle par le choix des termes définissant le type d'information contenue (« déclaration à la préfecture », « But », « Siège social »), l'ordre d'enchaînement normalisé des informations (date, lieu de déclaration, titre, objet, siège social pour une création) et la démarcation de ces informations, qui repose sur l'emploi d'un signe de ponctuation (le point).

Ces éléments sont donc déterminants pour traduire explicitement, en langage de programmation, la nature et l'agencement des informations, ie pour qu'il soit manipulable par des machines. Le point jouera le rôle que jouent souvent les tabulations ou autres virgules lors d'échanges de fichiers structurés : celle de délimitation des champs. La reconnaissance de pattern[148] intervient lorsque le point ne suffit plus à distinguer les types de rubrique. Ainsi, les expressions régulièrement utilisées telles que « change son titre, qui devient » ou « transfère son siège social » sont repérées en tant qu'indicateur de l'introduction d'un nouveau champ (soit, respectivement, « Nouveau titre » et « Nouveau siège social »).

Le retour à la ligne constitue, quant à lui, l'indice d'un nouvel enregistrement.

[148] Un pattern est « un phénomène ou une organisation que l'on peut observer de façon répétée lors de l'étude de certains sujets, auquel il peut conférer des propriétés caractéristiques. Un pattern constitue donc une solution générique à un type de problème fréquemment rencontré, en décrivant et formalisant les concepts sous-jacents à cette solution. », Page Wikipédia, https://fr.wikipedia.org/wiki/Pattern. Il est ici entendu au sens de chaînes de caractères apparaissant conjointement et régulièrement rencontrées dans le corpus. Intégrées dans un script de structuration, elles servent à identifier un type de données et à les délimiter.

Les scripts de structuration encodent de façon formelle l'ensemble de ces éléments pour fournir en sortie des fichiers de données structurées, manipulables par des machines et des applications numériques.
Il faut néanmoins ajouter que les termes « segmenteurs » (motifs récurrents qui permettent de délimiter une phrase) ainsi que les signes de ponctuation ne sont pas toujours présents dans les textes, en raison de leur non reconnaissance lors du processus océrisation. Il est alors possible d'utiliser Calliope, dans une finalité autre que celle de l'analyse lexicométrique : son module « Annotate » permet en effet d'extraire toutes les graphies possibles (dont les erreurs d'océrisation). Le corpus peut alors être analysé pour décider les expressions régulières à mettre en place voire, le cas échéant, constater que cela n'est pas envisageable – par exemple lorsqu'un même terme présente trop de variations de graphie qu'une expression régulière ne peut saisir.

Homogénéisation de la structure d'un point de vue diachronique
A cette première difficulté (traduire en langage informatique une structure implicite), s'ajoute une seconde. Loin d'être immuable, la structure des données du JO évolue en effet dans le temps, reflet de contraintes et d'objectifs différents. Ainsi, la structuration à mettre en place ne saurait se limiter à l'examen d'un seul exemplaire du JO mais doit prendre en compte l'ensemble des modifications que la structure a connues au fil des ans. Mathilde de Saint-Léger, qui a procédé à une analyse des fichiers, constate ainsi que « le format des fichiers est loin d'être unique. Par exemple, en 1960, le département est indiqué dans l'adresse de la déclaration tandis qu'au milieu de l'année 1963, le nom du département est indiqué en tête de toutes les déclarations du département puis, au cours de l'année 1966, le numéro du département est ajouté ; par ailleurs, les créations ou les dissolutions sont d'abord intégrées dans les annonces pour être ensuite regroupées à part, etc. ».
De même, le nombre de champs varie dans le temps : de 11 champs « reconstruits » en 1984, nous passons à 19 en 1997. Mais, si certaines informations contenues dans les rubriques de 1997 ne sont pas nécessaires à notre étude (comme le lien http vers le fichier en ligne), d'autres, présentes en 1984, n'apparaissent malheureusement plus en 1997. Ainsi, le changement de siège social, explicitement formulé en 1984 (par la formule « transfère son siège de … à … »), n'est plus que suggéré via la valeur « 2 » de la rubrique « Type » qui définit l'ensemble des modifications affectant les associations. Ces modifications peuvent concerner aussi bien une modification de titre, d'objet ou de siège social. Comme décrit dans la partie « Constitution du corpus », nous avons procédé à un filtrage des informations pour ne pas conserver inutilement les changements de siège social.
En ce qui concerne les indications géographiques (lieu de déclaration en préfecture ou sous-préfecture, adresse de l'association, etc.), les fichiers antérieurs à 1997 ont fait l'objet d'une précision : les régions ont été ainsi ajoutées au moyen de scripts – ce qui permettra de pouvoir procéder à des analyses de la distribution spatiale des associations à différentes échelles (ville, département, région). Tel n'est pas encore le cas pour les fichiers après 1997 (qui devront donc être complétés).
Enfin, la catégorisation des secteurs d'activité varie, à la fois dans le temps (modification du référentiel thématique en 2009 lors de la mise en place du Répertoire nationale des Associations – RNA[149]) et dans l'espace (selon le service préfectoral concerné, voire le(s)

[149] « Dans le cadre de la simplification administrative, un registre public et informatisé des associations (RNA – Répertoire national des Association) a été mis en place en 2009, qui permet :
- de mettre à disposition des services de l'Etat (administrations centrales et services extérieurs) les informations sur l'état civil des associations déclarées relevant de la loi du 1er juillet 1901,
- d'échanger automatiquement des données avec le *Journal officiel*,
- de fournir, conformément à la demande du CNVA (Conseil national de la vie associative), des statistiques fiables sur le milieu associatif, ce qui suppose la mise au point de nomenclatures adaptées : une nomenclature nationale d'objet social a été élaborée. », Site Associations.gouv : http://www.associations.gouv.fr/32-le-rna.html
Les informations enregistrées au RNA étant plus complètes (contenant les informations nominatives de la déclaration, telles que nom, prénom, profession, domicile, nationalité et fonction dans l'association), que celles diffusées par la DILA (titre, objet, siège social de l'association et adresse de ses établissements, durée, nature juridique de l'association, code d'objet social), elles pourront faire l'objet

personnel(s) affecté(s) à cette tâche[150]). Les descripteurs qui indexent une association peuvent donc différer, lors de la publication d'un avis de modification, par exemple.
Ainsi, avant la mise en place du RNA en 2009, le thème 01 correspondait à « Anciens combattants », tandis que dans la grille thématique établie depuis 2009, il correspond à « Activités politiques » ; le code 02 correspond à « Animaux » (avant 2009) et à « Associations caritatives, humanitaires, aide au développement, développement du bénévolat » (après 2009).

3.1.3 Nettoyage et normalisation des données

A l'instar des étapes précédentes, le nettoyage et la normalisation des données constituent une étape importante, réclamant minutie et rigueur.

Correction des graphies défectueuses
L'océrisation de documents scannés, bien que permettant l'accès numérique à des documents originellement en format papier, suscite autant d'avantages que d'inconvénients. Ou plutôt, ses avantages doivent être relativisés à l'aune de ses inconvénients, lesquels sont proportionnels à la quantité et à la qualité de documents traités. Le type et la qualité de papier, le type d'encre, le niveau de contraste entre caractères et papier, le type de typographie, etc. constituent autant de sources potentielles d'erreurs lors de la reconnaissance des caractères. Ainsi, un « I » majuscule sera parfois reconnu comme un « 1 ». Dans nos fichiers, le terme « œuvre(r) », par exemple, a donné lieu à différentes interprétations : « pceuvre » / « ceuvrer » et « ceuvre » / « liceuvre » / « izeuvre » / « luvrer » / « ceu vrer ». Ou encore, l'un de nos mots-clés « Anciens combattants » a donné lieu aux graphies suivantes : « anciens tombai ta nts » / » anciens combattonts » / « anciens combat tants » ; la ville de Mantes-la-Jolie est reconnue comme « mantes la jdlie », les chiffres romains sont transformés en lettres minuscules, etc. Ces erreurs mettent en échec l'analyse des données textuelles par des logiciels de textométrie

Afin d'y remédier, des scripts basés sur le recours à des expressions peuvent être développés. Mais il faut noter que certains scripts peuvent, à leur tour, générer des erreurs. Ainsi, en paramétrant le nombre de caractères à reconnaître dans une chaîne d'une certaine longueur, le script peut retourner comme résultats, « Aube » et « Aude » (département) ou « Sens » » et « Lens » (ville), qui ne diffèrent que d'une seule lettre. Ces termes ont en effet la particularité de partager, à la fois, le même nombre de caractères encodant les mêmes lettres, à une exception près. Cette exception explique l'erreur produite. Il est néanmoins possible d'intégrer dans le script un lien de conditionnalité entre la chaîne de caractères et le code postal, apparaissant dans un autre champ, de façon à limiter les interprétations « aberrantes », ie non conformes d'un point de vue sémantique. Il est également possible procéder à une vérification manuelle des résultats, une fois le script appliqué.

Nous voudrions surtout alerter ici sur le fait que, malgré la vigilance apportée à la correction de graphies défectueuses, il s'avère impossible de vouloir en traiter la totalité. Une certaine marge d'erreurs et d'inexactitudes doit être acceptée – à la condition que cette marge ne soit pas telle qu'elle oblitèrerait la qualité globale des documents. Il s'agit surtout d'en tenir compte lors de l'interprétation des résultats.

---

d'une analyse prosopographique, dans le cadre d'un prolongement du projet de recherche. Ces données permettront d'analyser la trajectoire des militants associatifs, en lien avec les questions mémorielles.

[150] Nous pouvons nous demander s'il existe des guides ou autres manuels visant à normaliser et à harmoniser, dans le temps et l'espace, les critères d'attribution de telle ou telle thématique à telle association. Que se passe-t-il lors d'un changement de personnel ? Comment s'effectue le suivi dans le temps d'une telle indexation ?

Transformations des caractères
Le logiciel Alceste nécessite un certain travail de mise en forme des documents avant de pouvoir être appliqué à un corpus. Cette mise en forme élimine les ambiguïtés sur lesquelles Alceste pourrait achopper.
Les mots en majuscule doivent ainsi être transformés en minuscule[151], les traits d'union en tiret bas (underscore), les points supprimés des acronymes ainsi que les espaces blancs inutiles[152], etc.
L'ensemble de ces opérations, aisées en soi, est néanmoins chronophage. Heureusement, des fonctionnalités d'OpenRefine (« Trim leading and trailing whitespaces » ; « Collapse consecutive whitespaces » ; « Transform to lowercase » ; etc.) et la définition de scripts basés sur des expressions régulières facilitent grandement cette tâche.

Typage des données
Le typage des données consiste à définir le type de valeurs que peut prendre une donnée, comme par exemple cadrer le format d'un champ « date » en JJ/MM/AAAA ou le format du code postal en 5 chiffres.
Cette action ne représente pas de difficultés particulières mais réclame une certaine vigilance, surtout lorsque le changement de format des documents est nombreux et régulier (passage d'un format XML à Excel, puis ouvert sous OpenRefine et finalement exporté en format de texte brut (.txt). Le typage des données peut être incorrectement interprété d'un format à un autre – ce qui peut occasionner des pertes d'informations. Ainsi, sur le fichier de l'année 2000, la date « 01/12/1999 » a été transformée, au gré des changements de formats de documents, en « 36495 ».
Afin d'homogénéiser les formats, nous avons donc veillé, à chaque transformation de format, à ce que le type de données soit défini et avons eu recours, lorsque nécessaire, à des fonctionnalités ou à des expressions régulières d'OpenRefine. Or, si OpenRefine reconnaît bien le format date, il lui applique par défaut le type suivant « 1999-12-01T00:00:00Z », dont le degré de granularité (indication de l'heure) ne nous est pas utile. Afin de le transformer en « 01_12_1999 », nous avons appliqué sur le champ date l'expression suivante :
value.toDate('yy-MM-ddT00:00:00Z','MMM-yy').toString('dd_MM_yyyy').

De même, nous avons intégré, pour le code postal et le numéro de département, un zéro supplémentaire aux numéros existants (codes postaux et départements de l'Ain à l'Ariège) via le script suivant : "00"[0,2-length(value)] + value ou "00"[0,5-length(value)] + value.

Par ailleurs, l'utilisation d'Alceste implique de convertir les informations contenues dans certains champs (dates, lieu de déclaration, code postal ou encore département) en variables illustratives. Afin de respecter les règles de formation de ces variables (variables désignées par le vocable « mots étoilés », dans le lexique alcestien), nous avons défini un script, dans OpenRefine, qui concatène l'intitulé de la colonne aux différentes valeurs du champ et lui ajoute une *.
Voici un exemple valable pour la variable « Département » :

1) Créer une nouvelle colonne « dpt2 » basée sur la colonne « dpt » :
   "add column based on" "dpt", remplie avec l'expression : "*dpt_"
2) Puis, concaténer le contenu de la nouvelle colonne avec celle de l'ancienne colonne, dans "Edit cell", à l'aide de l'expression suivante :
   « cells["dpt2"].value + cells["dpt"].value, if(isNull(),,)

[151] Plus précisément, Alceste exclut du traitement et de l'analyse linguistique les termes en capitales. Cette particularité peut être intéressante à exploiter. Par exemple, les termes dont la fréquence est très/ trop importante peuvent ainsi être « éliminés » du corpus par leur notation en majuscules. De cette façon, il est possible de procéder à deux traitements, avec et sans les termes « polluants » et de comparer leur distribution lexicale.
[152] Ces espaces blancs, interprétées comme des caractères par les machines, peuvent pervertir les statistiques.

3) Supprimer la colonne inutile ("remove this column") et renommer la nouvelle colonne générée par "rename column".

Homogénéisation des désignations

La gestion de données, notamment sur la longue durée, génèrent, outre les problèmes de structuration rencontrés auparavant, des variations dans la manière de désigner une entité. Ces variations peuvent provenir d'erreurs de saisie, de changement de « scripteur », de l'absence de référentiels terminologiques, de modifications de graphies, etc.

La normalisation des données doit donc tenir compte du problème que peut représenter la non homogénéité des dénominations. Le fait de désigner de façon unique une seule et même entité est particulièrement important dans le cadre d'un traitement lexicométrique, où une diversité d'appellations peut venir falsifier les statistiques lexicales. L'analyse de données textuelles, qui repose sur des principes statistiques de dénombrement des mots, utilise en effet, lors de l'élaboration des classes, des algorithmes d'appariement de chaînes de caractères. Si ces chaînes diffèrent d'un seul caractère, elles ne seront pas considérées comme identiques et donc comptabilisées deux fois. Ainsi, « From the point of view of theses algorithms, 'Post impressionnism', 'post impressionism' and 'post-impressionnism' are three different realities, as there is no complet match on the string level [96, HOOLAND et VERBORGH] ».

Seule une parfaite homogénéité des appellations est à même de garantir la qualité des analyses futures. Il faut donc veiller à ce que les différentes graphies d'une entité soient normalisées.

Open Refine dispose, fort heureusement, de plusieurs méthodes de clusterisation[153], fondées sur des mesures de similarités, qui permettent aisément de regrouper les différentes désignations. Une fois le repérage des graphies proches effectué, Open Refine propose de les fusionner, en fonction du degré de proximité qu'elles partagent.

Exemple : Liste des propositions de regroupement des lieux de déclaration – OpenRefine

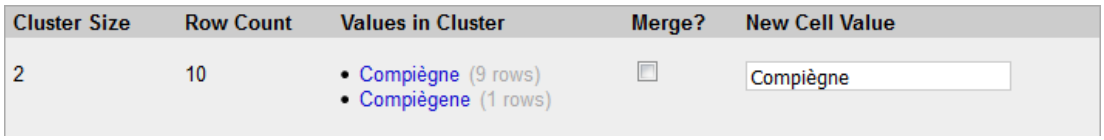
Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
3	6126	• Déclaration à la préfecture de police (6122 rows) • Déclaration à la préfecture de Police (3 rows) • Déclaration à la préfecture de police (1 rows)	☐	Déclaration à la préfecture de
3	1277	• Déclaration à la préfecture de la Loire-Atlantique (1275 rows) • Déclaration à la préfecture de la Loire-Atlantique (1 rows) • Déclaration à la préfecture de la Loire-Atlantique (1 rows)	☐	Déclaration à la préfecture de
2	144	• Déclaration à la préfecture de la Corse-du-Sud (143 rows) • Déclaration à la préfecture de Corse-du-Sud (1 rows)	☐	Déclaration à la préfecture de
2	91	• Déclaration à la sous-préfecture de Verdun (89 rows) • Déclaration à la sous-préfecture de Verdun (2 rows)	☐	Déclaration à la sous-préfectu

[153] Pour en savoir plus sur les méthodes de clusterisation d'OpenRefine, voir https://github.com/OpenRefine/OpenRefine/wiki/Clustering-In-Depth ou Linked Data (p. 104) ou Verborgh et De Wilde, Using Open Refine, Pakt Publishing, 2013.

Il est important de souligner qu'OpenRefine propose[154] mais ne procède jamais directement à la fusion des désignations. En effet, en fonction du type de méthode de clusterisation utilisé (« Key collision » ; « Nearest neighbor »), les résultats peuvent ne pas être pertinents, voire erronés. Il revient donc à l'utilisateur de valider les associations proposées.

Ainsi, en utilisant l'algorithme Nearest neighbor, dont le fonctionnement est plus agressif que celui de Key collision, OpenRefine propose de regrouper « Aube » et « Aude ».

Par contre, cet algorithme étant moins limité en termes de proximité que celui de Key collision, il peut proposer des clusters non identifiés par Key collision et qui peuvent s'avérer pertinents. Ci-dessous, la mauvaise saisie du nom de la ville de Compiègne a été reconnue par l'algorithme Nearest neighbor et associée à une autre graphie, elle, correcte.



L'ensemble de ces opérations de prétraitements visent à contrôler la cohérence des données et à les préparer aux étapes de traitement des outils d'analyse textuelle.
Malgré l'apparente simplicité de la description que nous venons de donner, les difficultés qui jalonnent la normalisation et le nettoyage de données sont multiples.
Le recours à des méthodes automatisées, via notamment des scripts en langage de programmation ou des outils tels qu'OpenRefine, si elles facilitent le prétraitement, butent parfois sur des « aberrations » qui sont indétectables d'une manière automatique. Ce qui ne relève pas d'un pattern, d'un schéma récurrent – autrement dit, l'insolite, le singulier – ne peut pas toujours être appréhendé ni traité automatiquement. Des vérifications manuelles s'imposent dès lors – ce qui nécessite d'y consacrer du temps…

3.1.4 De la normalisation à la modification des données ?

La frontière peut sembler floue, parfois, entre une normalisation des données et leur modification en vue de leur traitement. Si l'une est effectuée pour adapter les données aux formats requis par les techniques informatiques et statistiques – sorte d'adéquation formelle –, l'autre correspond davantage à une intervention sur le fond. Et, bien que proscrite en principe car pouvant s'apparenter à une dénaturation des données, concrètement, le prolongement de l'une par l'autre peut survenir. Ce fut le cas dans notre travail.
Nous avions constaté que certains lexiques étaient davantage représentés que d'autres, au point même de paraître surreprésentés au sein du corpus. Il en est ainsi de toutes les sections locales de fédérations nationales d'anciens combattants, telles que l'Union nationale des combattants (UNC). Lors de la déclaration, ces sections locales se réfèrent à l'intitulé et aux missions définies au niveau national par la fédération à laquelle elles appartiennent. D'où une présence plus importante, voire prédominante, de leur vocabulaire relativement à des associations isolées. Le contenu textuel de certaines associations, par la répétition des mêmes termes, des mêmes fragments, « polluait » en quelque sorte l'intégralité du lexique.

[154] La valeur proposée pour la normalisation de la graphie de l'entité est fonction de sa fréquence en termes d'occurrences.

Il ne s'agissait pas, bien évidemment, de nier le poids institutionnel de ces réseaux associatifs mais plutôt de pouvoir analyser finement le vocabulaire des déclarations sans que certains termes disparaissent des résultats, en raison de leur trop faible présence. En fait, deux logiques contradictoires s'affrontaient : comment concilier et rendre compte, à la fois, de la force réelle et dominante d'un certain type d'associations – dont le poids lexical en constitue le reflet – et de nuances lexicales plus fines, moins perceptibles, dont l'étude permettrait de saisir les multiples variations par lesquelles s'exprime la question mémorielle ? Le problème était d'autant plus déterminant que l'hypothèse de recherche initiale consistait non seulement à analyser les glissements lexicaux qui se produisaient au sein des déclarations d'anciens combattants mais également ceux d'associations relevant de « revendications particularistes ». Si celles-ci devenaient moins visibles, comment comparer les deux termes supposés de l'alternative mémorielle ?

Dans l'optique d'un traitement « égalitaire » de ces lexiques, différentes options ont été examinées.
- Devions-nous procéder à plusieurs traitements différents ? Le premier concernerait l'ensemble des déclarations – de manière à conserver le poids lexical et institutionnel réel des fédérations associatives ; les autres distingueraient deux sous-corpus, dont l'un analyserait le vocabulaire des associations affiliées à une structure nationale et l'autre s'attacherait à l'étude des déclarations hors fédérations – afin d'en saisir finement le vocabulaire, sans voir leur présence (en termes de fréquence et de cooccurrences) minorée au profit de « poids lourds » associatifs. Malheureusement, outre le problème crucial du temps de traitement, cette option n'a pas pu être appliquée pour des questions d'effectifs trop réduits. Le nombre des déclarations d'associations d'anciens combattants relevant d'organismes nationaux est tel dans notre corpus que le traitement des autres déclarations de façon distincte n'a pas été possible, d'un point de vue statistique. Ainsi, les diverses variantes locales de l'Union nationale des combattants représentent, en 2000, 465 déclarations (sur un total de 1 107 déclarations), sans compter les autres regroupements fédératifs (Fnaca, Association des anciens combattants, etc.)

- Devions-nous alors réduire l'ensemble de ces sections locales à un seul exemplaire ? Si le lexique employé au sein des déclarations des sections locales est identique, nous pourrions en réduire le poids par sa compilation au sein d'un exemplaire unique des différentes fédérations, une sorte d'exemplaire type.
Si cette solution peut paraître farfelue, voire dangereuse, elle s'appuie néanmoins sur l'hypothèse suivante : ces implantations locales proviendraient d'une décision prise au niveau national dans la mesure où, comme le montre les pics du graphique ci-dessous (réalisé par Mathilde de Saint-Léger), elles ont lieu à la fois massivement et par vague : un département connaît une année une vague de déclarations relevant de la même fédération (ici, l'UNC), suivi d'un autre département une autre année, etc.

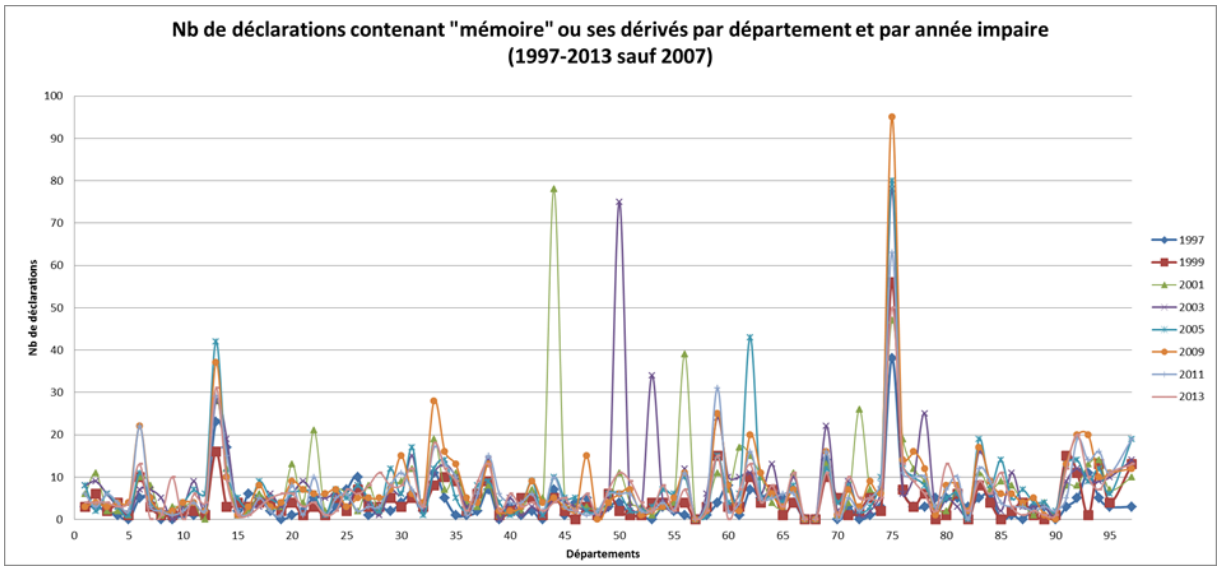


Figure 21 – Répartition temporelle et géographique des déclarations contenant « Mémoire »

Ces opérations d'implantations se déclinent comme suit : en 2001, pour les départements 22, 44, 56 (qui correspondent aux pics du graphique), elles concernent principalement l'Union nationale des combattants, qui représentent 114 déclarations distinctes sur 138 ; en 2003, dans le département 50, ce sont 65 déclarations sur 75.

Dans l'optique d'une réduction de ces associations, un premier problème, évident, surgit : cette manière de procéder supprime la visibilité de l'activité associative au niveau local ainsi que le poids de ces sections au sein de l'ensemble des associations. Le fait qu'une association nationale soit localement représentée est un élément d'information important, qui permet de mesurer l'importance institutionnelle d'un mouvement associatif mais aussi d'évaluer son pouvoir d'attraction au niveau régional. L'affiliation à une fédération révèle en effet la valeur accordée à celle-ci par les groupes implantés localement. Par ailleurs, ces associations locales adhèrent librement[155] à telle fédération nationale plutôt qu'à une autre. Les réduire à un exemplaire type reviendrait à nier cette décision.
S'ajoute à ce problème une autre difficulté d'ordre pratique : l'identité du contenu des déclarations n'est qu'apparente. D'infimes variations lexicales les caractérisent, repérées grâce à la fonction Clusterisation d'OpenRefine. Ainsi, sur les 465 déclarations relevant de l'Union nationale des combattants, nous avons dénombré de multiples variantes, dont en voici deux extraits [156] :
- 25 déclarations dont l'objet est : "maintenir, dans l'intérêt du pays, les liens de camaraderie, d'amitié et de solidarité qui existent entre ceux qui ont participé à la défense de la patrie ; défendre les intérêts moraux, sociaux et matériels de ses adhérents et de leurs ayants droit ; perpétuer le souvenir des combattants morts pour la France ; entretenir et développer des relations fraternelles entre les anciens combattants des nations amies ou alliées" ;
- 24 déclarations dont l'objet est : "maintenir les liens de bonne camaraderie créés entre les combattants de toutes les générations ; défendre les intérêts moraux, sociaux et

[155] Nous ne prenons pas en compte, dans le cadre de notre travail, les possibles facteurs politico-institutionnels présidant au développement d'une fédération sur l'ensemble du territoire. Justifier l'hypothèse d'une décision fédérative nationale gouvernant la création de sections locales demanderait une analyse fine et contextualisée qui n'est pas notre objet. C'est l'une des raisons expliquant l'abandon de cette solution.
[156] Nous ne donnons ici qu'un aperçu des multiples différences affectant aussi bien le titre (UNC ou Union nationale des combattants) que l'objet des déclarations. Les mêmes variations lexicales se retrouvent dans les déclarations des sections relevant d'autres fédérations nationales (Fnaca, Association des anciens combattants, etc.). Pour une présentation détaillée de ces variantes, voir en annexe (p. 230).

matériels de ces combattants ; perpétuer le souvenir des combattants morts pour la patrie et servir leur mémoire".

La diversité de contenu, infime parfois mais réelle, a conduit à devoir repenser l'hypothèse initiale. Les deux autres choix possibles consistaient soit à créer un contenu "hybride" mêlant les termes communs à l'ensemble des déclarations et ceux, « homogénéisés » (sic !), des termes différents ; soit, à reprendre la contenu de la première déclaration fondatrice de l'association, comme maître étalon. Si les déclarations locales se rapportent aux missions définies nationalement, ce texte pourrait alors devenir la référence. Et choisir le document d'origine comme référence permettait de ne pas avoir à sélectionner parmi les différentes versions qui se sont succédé. Or, dans le cas de l'UNC, cette association ayant été créée en 1918 (déclaration du JO du 11 décembre 1918), la mission de cette association s'avère déterminée, de façon évidente, par son époque et ne saurait correspondre aux finalités de sections locales beaucoup plus récemment créées. Reprendre ce document consisterait à dénaturer le contenu actuel par une intervention arbitraire que rien ne justifie
Précisons néanmoins que ces alternatives n'ont jamais été considérée comme sérieusement envisageables. Elles ne sont ici relatées que dans un but pédagogique : montrer les extrémités auxquelles peuvent conduire la situation problématique de données qui ne se laissent jamais aisément ni intégralement manipuler.

Le choix finalement retenu a consisté à ne pas "aplatir" le contenu des déclarations mais à analyser, à la fois conjointement et séparément, les déclarations contenant les mots-clés "mémoire" et "anciens combattants". Mais là encore, comme nous l'avons vu plus haut, les données se sont révélées rebelles…
Nous avons donc traité les fichiers conjointement mais en restant attentif, lors de l'exploration des résultats, aux résultats marginaux, peu visibles. La fonction de Calliope qui repère les termes émergents a facilité notre travail.

4 Traitements et production des résultats

Dans cette partie, nous ne décrirons pas de façon détaillée l'ensemble des traitements opérés par les deux outils d'analyse textuelle sélectionnés, Calliope et Alceste, ni les méthodes implémentées. Une telle présentation présenterait en effet le double inconvénient d'être à la fois fastidieuse et inutile. Nous nous attacherons plutôt à mettre en relief, dans une perspective comparative, les points saillants des deux outils, notamment lors de la phase, essentielle, d'élaboration des lexiques. Précisons que cette partie n'aborde la production de résultats que par les éventuels biais que les principes de fonctionnement de l'outil peuvent y introduire. La phase d'interprétation des résultats et leur comparaison feront l'objet de la troisième partie de ce mémoire.

4.1 Extraction et élaboration du lexique

Chaque outil d'analyse de données textuelles possède ses propres méthodes d'extraction et d'élaboration du lexique, sur lequel seront ensuite appliquées les mesures statistiques (comme le calcul de la distance entre termes) donnant lieu à la classification des éléments.
Cette phase de l'analyse demande, tout comme celle de constitution du corpus, vigilance et itérativité. D'elle dépend, notamment, la qualité des résultats, leur pertinence au regard de la finalité recherchée. Nous allons montrer dans les paragraphes qui suivent certaines des questions que soulèvent cette étape et les précautions qu'elles nécessitent, que nous illustreront par des exemples issus des deux outils utilisés, Calliope et Alceste.

4.1.1 Calliope : une élaboration du lexique supervisée et itérative

La méthode d'élaboration du lexique sous Calliope correspond, selon nous, à une approche supervisée en ce sens qu'elle implique une intervention humaine dans son processus[157]. Contrairement à Alceste[158], dont le fonctionnement est purement statistique, Calliope mêle procédures quantitative et qualitative. Ce faisant, il constitue, à nos yeux, un mixte entre les CAQDAS et les outils de statistique lexicale proprement dits.
Le traitement de données textuelles par le logiciel Calliope nécessite en effet, au moment de l'élaboration du lexique, une supervision humaine pour valider le vocabulaire brut extrait, selon un seuil de fréquence à fixer en fonction du type de corpus et de la finalité recherchée. Cette validation est le fait d'un expert du domaine qui, seul, peut déterminer la pertinence des choix à effectuer, en fonction de ses hypothèses initiales. L'intérêt d'une telle intervention est de permettre à l'expert, au chercheur, de participer au choix de ce qui doit être analysé par la machine et d'éventuellement de corriger, de mettre en valeur ce qui pourrait autrement être nivelé par un traitement statistique non supervisé.
Cette phase est donc essentielle car, comme le précise le manuel d'utilisation de Calliope, elle « consiste à sélectionner les termes [jugés] pertinents et représentatifs », « aptes à en symboliser le contenu sans ambiguïté »[159]. Toute la difficulté réside précisément dans cette estimation du degré de pertinence et de représentativité des éléments, en vue
La méthode consiste, une fois l'extraction brute du vocabulaire réalisée, à distinguer manuellement les termes extraits en deux groupes : d'un côté, les termes que l'on qualifie de « validés » (ie, ceux retenus pour indexer les documents du corpus) et de l'autre, les

[157] Cette approche supervisée est à distinguer de ce que l'on nomme « apprentissage supervisé » en intelligence artificielle, qui est une technique d'apprentissage automatique qui cherche à produire automatiquement de nouvelles règles à partir d'un corpus d'apprentissage.
[158] Ces deux outils offrent la possibilité de modifier leurs divers paramètres, les dictionnaires intégrés, etc. Un certain type d'intervention humaine a donc lieu dans les deux cas.
[159] Voir notamment la présentation en ligne : http://www.calliope-textmining.com/aide/fr/059.html

termes désignés comme « préférentiels » et « synonymes »[160]. Ce regroupement peut être délicat à effectuer, et demander à être revu, de manière itérative. L'évaluation de la pertinence du choix lors de la validation des termes implique en effet de nombreux questionnements devant guider le processus. Chaque action de validation nécessite en effet de se demander quels peuvent en être les éventuelles conséquences sur les résultats produits : Quels termes retenir ? Quels termes exclure ? Quels termes préférentiels sélectionner ? Qu'appelle-t-on un synonyme ? Comment traiter la polysémie des termes ? Dans la mesure où ces choix influencent les résultats finaux, une vigilance particulière doit leur accordée. Pour illustrer ces difficultés, nous sélectionnerons quelques exemples parmi notre corpus.

Comment exprimer la granularité territoriale des activités associatives ?

L'une des finalités du projet dans lequel nous nous inscrivons est de cerner au plus près la répartition spatiale des associations et de leurs activités autour de la question mémorielle : cela signifie non seulement analyser la distribution spatiale des associations en fonction de leur lieu de déclaration (pour vérifier si des corrélations se dessinent entre un type d'association et un territoire) mais aussi tenter de voir, au sein de leur vocabulaire même, si les différentes échelles territoriales auxquelles elles se destinent apparaissent. Une association se veut-elle d'envergure nationale (voire internationale), régionale, locale ? Leurs actions ont-elles vocation à agir au niveau local ? Si tel est le cas, est-ce au niveau d'une commune, d'un village, d'un arrondissement d'une ville, d'un quartier ? Ou ces actions se déploient-elles à l'échelle nationale, voire internationale ?

Qu'en dit leur lexique ? Après analyse, il s'avère que celui-ci comporte de nombreux termes désignant différents niveaux territoriaux. Afin de conserver une certaine finesse dans la granularité spatiale exprimée, il a été décidé de répartir les termes selon les catégories de l'administration territoriale française et des divers modes d'habitat. L'élaboration de ce vocabulaire dédié aux échelles « géographico-administratives » devait ainsi permettre de pouvoir analyser la portée géographique des activités des associations. Les termes (ainsi que leurs formes fléchies) de « canton », « communal », « département », « région », « Nation » ont donc été validés en tant que tels. D'autres termes, exprimant d'autres découpages de l'espace (types d'agglomération, d'habitat, d'environnement), ont également été pris en compte : il s'agit notamment de « rural », d'« urbain », de « village » (contenant le terme « hameau »), de « ville », etc.

Par contre, les noms des départements tels que « Aisne », « Aube » ou encore « Ardennes », « Finistère » et « Var » ont été subsumés sous la catégorie « département ».

Ce choix s'explique par le fait qu'il s'agissait prioritairement de mettre en valeur, au moyen du lexique, le niveau territorial auquel s'exerçait l'activité associative et non la désignation du lieu lui-même, qui peut être retrouvé par un tric croisé sur la variable « Lieu de déclaration ». Or, la lecture des résultats nous a amenés à devoir revoir cette répartition. La visualisation du diagramme stratégique a effet mis en avant des résultats problématiques, ayant nécessité une remise en question des choix terminologiques précédemment effectués.

Le poids informationnel du terme préférentiel « département » était tel qu'il participait à la construction de l'un des clusters les plus fédérateurs du diagramme stratégique. Ce poids, construit, ne reflétait pas la « réalité textuelle » dans la mesure où seul le regroupement de différents termes en son sein l'expliquait. Il a donc été convenu par la suite de procéder au « dégroupement » de ces termes, afin de représenter davantage leur poids informationnel « réel » au sein du corpus. De même, pour les différentes régions françaises, qui ont été finalement exclues du mot-maître « région » pour les mêmes raisons. A noter, à ce propos que deux régions (Bretagne, Corse) faisaient figure d'exception. En raison de leurs liens privilégiés (ou supputés) à la question mémorielle selon le discours général, elles avaient été dès le départ classées hors de « région ». La suppression du regroupement des régions en un seul méta-terme a donc permis de rétablir un traitement plus égalitaire entre les termes.

[160] Dans la terminologie de Calliope, les termes préférentiel désignent les termes (maîtres-mots) sous-lesquels seront rangés les synonymes, à la signification proche.

Comment exprimer le lien familial ?

En raison du lien au passé qui s'exprime dans le phénomène mémoriel, le vocabulaire des déclarations d'associations fait état d'une thématique généalogique importante. Afin de pouvoir appréhender finement le lien entre mémoire et filiation, nous avons choisi de valider les termes suivants : « fils », « fille » (sans les regrouper, pour ne pas occulter une éventuelle dimension du genre), « orphelin », « pupille », ainsi que le terme générique « génération », qui permettra d'indexer les syntagmes nominaux suivants : « génération future », « jeune génération ».

Cette répartition peut néanmoins poser problème en cas de polysémie Ainsi, un même terme, qui indique un rapport de filiation, peut exprimer un sens « réel » (les enfants de parents physiques) ou symbolique, figuré (les enfants de la Nation).

Le terme « ayants-droits » soulève d'autres difficultés. Les ayants-droits désignant l'ensemble des personnes détenant des droits en raison de leur lien familial au combattant décédé, leur mention apparait fréquemment au sein des déclarations d'anciens combattants, destinées, en partie, à protéger leurs intérêts moraux et matériels. Afin de représenter toutes les personnes concernées (ascendant, descendant et conjoint notamment) – dont les termes cooccurrent très fréquemment –, fallait-il les rassembler sous le seul terme « ayants-droits » ? Après discussion, le choix a été fait de procéder à ce regroupement. Or, les résultats montrent, qu'en 2010, ce terme participe à l'un des clusters les plus important du corpus : en aurait-il été de même si cette agrégation n'avant pas été effectuée ?...

Difficulté à gérer les termes polysémiques

D'autres difficultés concernent la gestion des termes polysémiques mais dont un sens principal domine au sein du corpus. Ces termes posent des problèmes de catégorisation. Il en est ainsi du terme « création », qui désigne le plus souvent dans les déclarations le processus de création artistique mais peut également se référer à l'acte par lequel est officiellement instaurée une institution, un organisme, etc. Etant donné la prédominance de son sens artistique, il a été décidé d'associer ce terme au méta-terme « Art et création », qui rassemble art(s), art(s) vivant(s), artiste(s), artistique et création(s). Il en est de même pour le terme « développement » rangé avec les termes relation et échange, sous celui d'« échange ».

Qu'en est-il de l'influence de ces choix sur les résultats ?

Importation d'un lexique validé sur un autre corpus : **une adaptation nécessaire**

Un lexique déjà validé dans le cadre d'un autre corpus mais se rapportant aux mêmes thématiques (ici, question mémorielle) et aux mêmes acteurs (ici, associations) peut être intégré afin de faciliter et de limiter le temps dédié à l'étape de validation. Tel a été le cas pour notre corpus de déclarations au JO, dans lequel nous avons inséré le vocabulaire validé issu de fichiers de la Mairie de Paris. Si de nombreux termes sont communs aux deux lexiques, certains arrangements ont dû être faits, afin d'adapter un vocabulaire aux particularités lexicales de l'autre. Par exemple, le méta-terme « Antillais » issu du fichier de la Mairie de Paris a été transformé en « DOM TOM » afin de pouvoir subsumer Guadeloupe et Réunion, qui ne figuraient pas dans le corpus Mairie de Paris. Cette appellation plus large permet d'indexer un ensemble de notions relevant de la même catégorie Par ailleurs, un certain nombre de nom de rues parisiennes, naturellement présent au sein de fichiers issus de cette ville, n'a été retenu pour le corpus JO en raison de leur absence.

Validation de lexiques volumineux : une tâche rigoureuse, à maintenir dans le temps

Enfin, dernière difficulté, la taille du lexique à valider. Lorsque celle-ci est importante (de l'ordre de plusieurs milliers de termes dans le cadre de notre corpus), l'étape de validation peut vite devenir fastidieuse. C'est l'une des raisons, évoquée ci-dessus, à l'intégration de lexiques déjà validés. Outre le temps requis, valider un lexique volumineux implique également, et surtout, de devoir conserver une logique identique dans le traitement des termes tout au long du processus. Si le choix a été fait de ne regrouper, sous un maître mot, que les synonymes stricts ou les formes fléchies d'un terme, il faudra appliquer cette décision pour chaque terme rencontré. Or, il est possible qu'après plusieurs heures de travail

(ou plusieurs séances s'échelonnant dans le temps), la rigueur de la méthode s'émousse, que les choix effectués précédemment s'estompent…

Nous espérons avoir réussi à montrer, à travers ces quelques exemples, que la tâche de validation du lexique dans Calliope ne doit pas être négligée. Elle constitue une étape essentielle, et comme telle, délicate à aborder. L'enjeu étant la qualité des résultats produits, procéder à une validation rigoureuse et pertinente requiert expérience et recul, discussions et allers-retours.
Pour conclure, précisons qu'une sélection du vocabulaire a été effectuée : les verbes ainsi que les adjectifs ont ainsi été exclus lors du traitement des données avec Calliope. En effet, dans une optique de représentation des thématiques du corpus, la présence de verbes au sein des clusters n'apportent pas d'information pertinente (tout au plus voit-on le verbe « perpétuer » associé au terme « souvenir »). Cette présence peut par ailleurs empêcher la prise en compte, lors de la classification, d'autres termes au poids informationnel (fréquence relative et typicalité) moins fort mais au poids informatif plus fort. Les adjectifs, de leur côté, s'avèrent trop polysémiques pour être utiles à l'analyse. Ainsi, l'adjectif « social » peut être rapporté à de nombreux substantifs (insertion, action, lien, intérêt, œuvre, justice, objet, équilibre, aide, initiative, etc.) sans qu'un dénominateur commun thématique puissent être trouvé. C'est pourquoi seuls les segments contenant l'adjectif social qui représentaient un intérêt pour l'analyse ont été conservés, lors de la validation du lexique, tels que « lien social », « action sociale » et « insertion sociale ».

4.1.2 Alceste : un processus d'élaboration du lexique non supervisé

Contrairement à Calliope, le processus d'extraction et de traitement du lexique dans Alceste ne demande pas d'intervention humaine. En ce sens, il est non supervisé : la machine extrait automatiquement de l'ensemble des documents les éléments lexicaux sur lesquels les calculs statistiques seront ensuite appliqués. Rappelons qu'il est néanmoins possible de modifier les paramètres de l'outil (réglage du seuil de fréquence, par exemple, ou de la taille de la fenêtre comprenant les unités de contexte) et les dictionnaires intégrés, en fonction, notamment du type de corpus et du degré de spécialisation d'un lexique (langue de spécialité, par exemple). Mais, à l'instar de Calliope, l'étape d'extraction et de catégorisation du vocabulaire ne va pas sans générer certaines difficultés. Nous voudrions insister ici sur le problème spécifique que représente le processus de de lemmatisation, au moment de l'identification et de la catégorisation des unités lexicales de base.

Lemmatisation et regroupement lexical : un processus à risque
Nous avons abordé les risques du processus de lemmatisation en première partie Rappelons simplement ici que la lemmatisation est un processus linguistique qui consiste à transformer les formes graphiques (ie, les mots tels qu'ils apparaissent au sein des corpus textuels) en leur forme canonique. Ainsi, une forme verbale est réduite à son infinitif, un substantif pluriel à son singulier, une forme élidée est reproduite sans élision (l' devient le). Au sein d'Alceste, cette opération de réduction - regroupement lexical est particulièrement puissante[161], ainsi que nous allons le montrer au travers de trois exemples.
Comme le montrent les images ci-dessous, trois groupes lexicaux ont été créées par Alceste : « Act+ion », « Nationa+l » et « Arme+ »[162], qui rassemblent l'ensemble des termes de même racine.

[161] Bien qu'elle puisse, comme n'importe quel processus, générer des erreurs. Ainsi, dans notre corpus, le nom propre « De Gaulle » a été assimilé au verbe « gauler ». Mais ces identifications défectueuses sont heureusement rares.
[162] Le signe « + » indique l'endroit où Alceste a effectué la césure des termes, séparant la racine de ses désinences.

Act+ion

2338	4	Y act+ion	acteur
1452	17	Y act+ion	acteurs
5159	10	Y act+ion	actif
2662	9	Y act+ion	actifs
77	183	Y act+ion	action
1284	172	Y act+ion	actions
1267	13	Y act+ion	active
1875	8	Y act+ion	activement
6132	1	Y act+ion	activistes
934	48	Y act+ion	activite
969	129	Y act+ion	activites

cette **action** s' inscrit dans le cadre d' une action
t dans le cadre d' une **action** nationale pilotee par les associations
destines a ce type d' **activites** ou en itinerantes; par l' organisat
ergenerationnel, cette **activite** s' inscrit dans le cadre des echange
s, locaux, exterieurs, **acteurs** du monde social et politique, d, deve
elopper un partenariat **actif** avec les collectivites territoriales, e
e, participer a toute **action** a dimension sociale, educative, econom
rancazal, haut lieu d' **activites** aeronautiques civiles et militaires
handicapes **activistes** solidarite ergotherapie, hase, col
ement durable dans ses **actions**, gerer l' espace memoire du patrimoin
n hummel france et les **actifs**; developper des activites entre retrai
actifs; developper des **activites** entre retraites de mann hummel fran
ement aider toutes les **actions**,

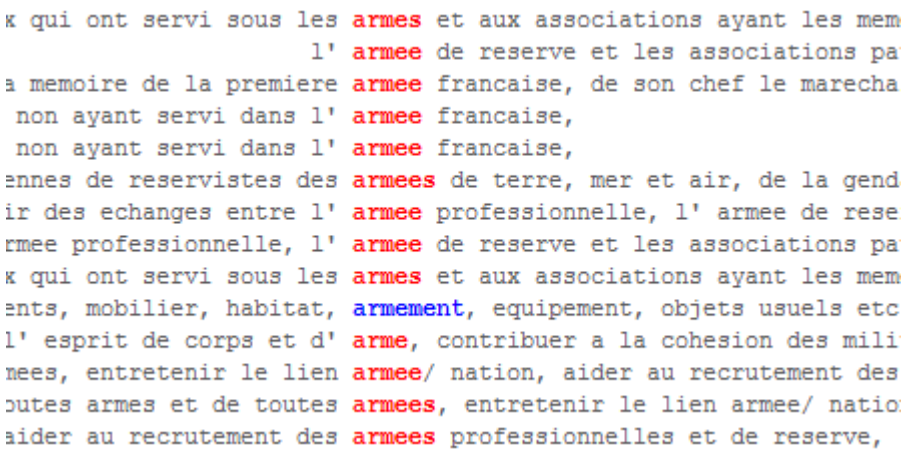
Nationa+l

436	54	Anationa+l	national
267	934	Anationa+l	nationale
3315	1	Anationa+l	nationalement
1562	43	Anationa+l	nationales
8624	4	Anationa+l	nationalite
7183	3	Anationa+l	nationalites
4341	1	Anationa+l	national'
4024	8	Anationa+l	nationaux

union **nationale** des officiers de reserve nice c
stoire a l' echelle locale et **nationale** et internationale; elle s' insc
t dans le cadre d' une action **nationale** pilotee par les associations na
pilotee par les associations **nationales** gens du voyage et un collectif
aires, les anciens du service **national** et les ressortissants de l' onac
gulierement sur le territoire **national**.
e monde, quelle que soit leur **nationalite**, afin que leur sacrifice pour
t des valeurs patriotiques et **nationales**.
de la paix et de la securite **nationale** et internationale et de defendr
bration des differentes fetes **nationales**, et la nouvelle annee traditio
journee **nationale** d' hommage des morts pour la fr
s manifestations patriotiques **nationales** en accord avec la municipalite
actions d' aides humanitaires **nationales** et internationales.

Arme+

1951	7	Aarme+	arme
523	55	Aarme+	armee
1258	22	Aarme+	armees
2954	2	Aarme+	armement
1364	38	Aarme+	armes



L'intérêt, voire la pertinence de ces affectations lexicales dépend, en fait, de nombreux facteurs (type de corpus, d'analyse et de résultats attendus). Qui veut étudier finement le vocabulaire de l'« identité nationale », par exemple, ne souhaitera pas voir associés les différents termes de la question. Quel lien unit une action déployée à l'échelle « nationale » et la « nationalité » d'une personne ? Par ailleurs, bien que le lien sémantique entre « arme », « armée » et « armement » paraisse évident, les réunir au sein du même méta-terme peut soulever questions et problèmes interprétatifs. Pour finir, le rapprochement d'« action » (ou « activité ») et d'« activiste » semble hasardeux, pour ne pas dire plus.
Nous constatons donc que le processus de lemmatisation, s'il peut s'avérer utile – en ce sens que différents termes éparpillés d'une même famille lexicale se verront réunis, et donc leur poids informationnel augmenter –, peut, à l'inverse, générer de nombreux problèmes. Les résultats en seront nécessairement influencés.
Notons, néanmoins, qu'une éventuelle modification des regroupements issus de la lemmatisation peut avoir lieu : les agrégats créés par l'opération de lemmatisation peuvent en effet être annulés, modifiés, revus. « On peut décider, par exemple, de supprimer le regroupement des formes « vert » et « verts » en considérant que l'usage de ce mot au singulier ou au pluriel n'a pas le même sens. C'est au fur et à mesure des utilisations et en fonction de la spécificité de chacun des corpus que l'on ressentira le besoin d'intervenir sur les catégories de lemmatisation d'Alceste. [36, GUERIN-PACE] »

Alors qu'avec Calliope, en raison de sa finalité, nous avions exclu les verbes et les adjectifs, ils ont été conservés sous Alceste. Cela constitue une autre différence entre les deux outils : la représentation thématique qu'opère Calliope n'a que peu d'intérêt à la conservation de ces types de termes tandis que les spécifications de types de discours élaborées par Alceste les rendent intéressantes. En effet, replacés dans leur contexte d'apparition, les verbes ou les adjectifs prennent tout leur sens : « honorer la mémoire des conscrits », « honorer nos morts », « honorer la mémoire des maquisards », « honorer les sacrifices », « honorer leurs souvenirs », etc. désignent des actions différentes. Et le fait de conserver les verbes permet, dans ce cas, de préciser les thématiques de Calliope (mort, souvenir, sacrifice, etc.). Nous verrons plus tard, lors de l'analyse des résultats, que les verbes peuvent, en outre, servir à caractériser une classe de discours : la classe « patrimoine » se verra ainsi spécifiée par une surreprésentation des verbes au sein de ses énoncés.

Pour conclure, il suffit de dire que les deux méthodes d'élaboration et de catégorisation du lexique comportent leurs propres avantages et limites : l'une et l'autre intervenant sur le vocabulaire (de manière supervisée ou automatique) à dénombrer et à comparer, leur influence sur les résultats est donc réelle. Le seul garant contre ces risques demeure la prudence de la démarche !

5 Apports et limites méthodologiques

Nous aimerions conclure cette partie par une présentation des limites et des apports méthodologiques de notre étude. Le projet dans lequel nous nous inscrivons consistait initialement à étudier l'évolution du vocabulaire mémoriel des déclarations d'associations au *Journal officiel*, de 1945 à nos jours. Pour cela, nous avons utilisé deux outils d'analyse de données textuelles, Calliope et Alceste. En raison de nombreuses contraintes, les résultats produits, ainsi que le périmètre du projet, ont dû être fortement revus à la baisse : ainsi, seules trois années (1984, 2000 et 2010) de déclarations ont pu être traitées, sans distinction de contenu (entre titre et objet des déclarations) ni de mot-clé.
Cette réduction drastique ôte toute représentativité au corpus tout en limitant fortement la portée des résultats. Mais ceux-ci, pour restreints qu'ils soient, avaient surtout une fonction exploratoire et ce, à un double niveau : une première exploration de données dont la résistance intrinsèque demandait à être testée et évaluée par une méthode qu'il convient d'affiner ; une première exploration d'un domaine de compétences réalisée dans le cadre d'une formation continue, accompagnée d'une mise en pratique lors d'un stage de 3 mois.
Nous présentons, brièvement, quelques-uns des facteurs qui ont contraint la portée de notre étude :

- **Une durée de stage limitée**
La courte durée du stage (3 mois) a fortement déterminé la taille du corpus finalement traité et les résultats produits. En effet, un projet d'analyse de données textuelle requiert du temps, qu'il est souvent (toujours ?) très difficile à estimer – et lorsque une estimation temporelle est faite, elle tend souvent à être sous-évaluée. L'accès aux données, leur lisibilité par la machine, leur homogénéité et leur normalisation sont des étapes chronophages – comme nous espérons l'avoir montré tout au long de cette partie La constitution d'un corpus est une opération délicate devant concilier contraintes internes, externes et adéquations aux objectifs du projet. Lorsque les contraintes sont incontournables, une révision du périmètre du projet doit alors avoir lieu.
Cette contrainte temporelle explique donc, pour partie, la réduction (temporaire) de la couverture du projet. De l'analyse des déclarations de 1945 à nos jours, le projet a été limité au seul examen de trois années, parmi l'ensemble des fichiers à notre disposition.

- **Une disponibilité problématique et une lisibilité défectueuses de certaines données**
Le manque de structuration des données antérieures à 1985 et leurs graphies défectueuses ont également pesé sur la constitution du corpus. Ces problèmes ont en effet empêché la sélection de données parmi un éventail plus large. Ayant fait l'objet de nettoyage, en parallèle de notre travail, elles sont désormais disponibles pour de futurs traitements et analyses. Par ailleurs, le format de certains fichiers (les microfiches archivées à la BnF) en interdit tout traitement. La période 1985-1996 ne peut donc actuellement être intégrée à notre projet.

- **Une période de découverte et d'apprentissage de l'analyse de données textuelles**
Ce travail s'inscrit dans le cadre d'une reconversion professionnelle qui correspond à une découverte du domaine de l'analyse de données textuelles. La masse d'éléments (tant théoriques que pratiques) à assimiler sur une courte période était telle qu'elle constitue sans doute l'un des facteurs essentiels à la limitation du projet, et aux possibles erreurs produites. Une longue phase d'apprentissage de connaissances et de compétences est en effet requise dans tout projet d'analyse de données textuelles et leurs outils. Assimiler les principes statistiques sur lesquels sont fondés les logiciels (dont nous avons vu qu'ils déterminaient fortement les résultats) tout en tentant d'en maîtriser les fonctionnalités est un processus lent, qui comprend non seulement une pratique distanciée des outils (lorsque possible…),

une exploration des multiples possibilités offertes, des tests de paramétrage[163], etc. mais également la lecture de la littérature sur le sujet pour en saisir les principaux concepts, de multiples échanges avec des collègues plus expérimentés, etc. Il s'agit d'être en mesure, après la production des résultats, d'en sélectionner certains pouvant présenter une pertinence pour l'analyse, de déceler parmi la profusion de restitutions celles qui offrent un potentiel interprétatif intéressant. S'ajoute à cela, de façon concomitante la découverte du projet de recherche, du domaine de la sociologie de l'action publique, des hypothèses, de la problématique du projet, etc. Cette connaissance est primordiale pour sélectionner les résultats pertinents à l'analyse.
La capacité à traiter les donnée, à lire et à choisir les « bons » graphiques, cette sensibilité aux résultats, demandent une prise de distance par rapport à l'objet qui doit toujours être guidée par la problématique. Ces compétences et connaissances ne peuvent s'acquérir que par un temps long d'apprentissage, composé d'essais et d'erreurs – ce que n'a que très imparfaitement permis le stage et qui reste encore largement à explorer !

Les écueils rencontrés ont néanmoins permis de confirmer l'intérêt de la démarche pour le projet. Outre l'accès « facilité » à une grande masse de données qu'une lecture « humaine », linéaire, aurait eu du mal à appréhender, elle a ouvert de futures pistes d'exploration et d'interprétation des données. Cette méthode demande bien sûr à être précisée, développée, modifiée mais elle pourra, pour ce faire, s'appuyer sur les briques déjà posées :

- **L'adéquation des outils lexicométriques sélectionnés au type de documents traités et de résultats attendus**
Le choix des outils retenus pour l'analyse des déclarations, Calliope et Alceste, s'est vu confirmé dans sa pertinence : leur méthode est en effet adaptée au type de matériau textuel constituant les déclarations. Leur contenu est, comme nous l'avons vu, formaté, descriptif, composé essentiellement de mots pleins (noms, verbes, adjectifs, adverbes) et convient parfaitement au postulat linguistique implémenté dans les outils. Si celui-ci peut constituer un risque dans le cadre d'une analyse de discours fine, il s'avère respecter, par ailleurs, tout à fait la finalité du projet, qui est l'analyse de l'évolution lexicale des déclarations au JO en ce qui concerne la question mémorielle.

- **L'intérêt de multiplier les outils et les démarches**
Bien que relevant de la même catégorie générale d'outils (logiciels d'analyse de données textuelles), Calliope et Alceste diffèrent notamment quant à la méthode de classification utilisée ou le processus de constitution du lexique. Cette différence méthodologique se traduit évidemment par une différence de résultats. Comme nous le verrons dans la partie suivante, l'intérêt de multiplier ces démarches d'exploration des données est de pouvoir en offrir des points de vue diversifiés qui, en se complétant, enrichissent l'analyse.

- **Une prudence acquise**
L'ensemble des points évoqués ci-dessus n'a d'autre fonction que d'alerter sur les risques inhérents à tout projet d'analyse de données textuelle. Car ce n'est qu'une fois ces risques identifiés et pris en compte que l'intérêt de ce type de démarche peut révéler toute l'ampleur de son potentiel.

Afin de résumer ce bilan d'ordre méthodologique, nous proposons un tableau présentant les éléments qui nous paraissent importants à identifier et à évaluer dans un projet d'analyse de données :

[163] La courte durée du stage ainsi que la découverte du domaine expliquent par ailleurs que nous n'ayons pas procédé à des variations du paramétrage des outils ni à des modifications de leurs dictionnaires internes, pour davantage vérifier l'adéquation (ou non) des résultats aux besoins de l'analyse.

Définition du périmètre du projet et choix d'outils	
Périmètre évolutif du projet : adapter les choix à la finalité du projet et aux ressources disponibles, de façon itérative	Quel est le corpus potentiel disponible ? Quelles sont les sources accessibles ? Combien de temps le traitement et l'analyse des résultats nécessitent-ils ? Y a-t-il adéquation, à chaque étape, entre les choix effectués et la finalité du projet ? Quelles sont les ressources (compétences, financières, temporelles) disponibles ? > Etablir un **plan de traitement et d'analyse** (estimation des étapes et des difficultés, évaluation du temps de pré-traitement et d'analyse, etc.) > **Principe itératif** : revoir le périmètre du projet en fonction de son état avancement, des difficultés qui se présentent et des ressources
Choix des outils d'analyse de données textuelles / Mode de traitement des données	Quels sont les principes statistiques sur lesquels sont fondés les outils (méthode de classification par exemple) ? : Quelles méthode de classification ? Quelles sont les unités textuelles de base ? Existe-t-il un processus de lemmatisation ? Qu'est-ce qui est extrait, compté, classé et de quelle manière ? > Le **choix de l'outils** repose sur son adéquation au type et au volume du corpus, d'analyse (de contenu vs. discours) et de problématique étudiée, de résultats attendus
	> La procédure de traitement est-elle intégralement automatique ou admet-elle une intervention humaine (hors paramétrage) ? : démarche supervisée ou non supervisée
	> Quels coûts (compétences, temps de (pré)-traitements, prix) impliquent le choix de tel outil ?
	> Quels sont les modes de restitution des résultats ? Des aides à l'interprétation (édition de rapports) sont-elles fournies ?
Constitution du corpus : précaution dans la sélection des données	
	> S'assurer de la **disponibilité des données** Sinon évaluer le coût (temps, budget) d'accès aux données et/ou modifier le périmètre du projet
	> S'assurer de la **lisibilité des données par la machine** Sinon évaluer le coût (compétences requises, temps, budget) de structuration et de nettoyage des données et/ou modifier le périmètre du projet
	> S'assurer de **l'homogénéité du corpus** : même source, mêmes conditions d'énonciation, même type de données textuelles, etc.
	> En fonction des contraintes de constitution du corpus, s'assurer de la **représentativité du corpus** Sinon sélectionner d'autres données et/ou modifier le périmètre du projet
	> Déterminer la **couverture temporelle** adéquate (en fonction des contraintes)

	Sinon modifier le périmètre du projet
	> S'assurer de la **taille suffisante du corpus** aux regard des postulats statistiques des outils Sinon procéder à des regroupements justifiés
Résultats : précaution dans l'interprétation des résultats	
	> Ne jamais oublier les **objectif du projet, qui doivent guider la sélection des résultats et leur interprétation**
	> Ne jamais oublier que les **résultats dépendent des conditions de production du corpus**
	> Ne pas considérer les **restitution graphiques** comme décrivant une réalité mais comme **un point de vue**, admettant plusieurs interprétations possibles en fonction de la finalité d'un projet (une aide au questionnement)
Intérêt d'appliquer de multiples outils	
	> Multiplier les méthodes : **diversité** des résultats
	> **Complémentarité** des méthodes et des résultats

Troisième partie
Présentation des résultats et analyse comparative

1 Première approche quantitative du phénomène mémoriel

Avant de présenter et de comparer les résultats obtenus via les deux outils d'analyse de données textuelles sélectionnés, Calliope et Alceste, nous allons procéder à une première analyse des thématiques, d'un point de vue quantitatif. Ces dénombrements ont pour but de fournir un premier aperçu du phénomène mémoriel.
La présentation des résultats produits à l'aide des deux logiciels d'analyse de données sera, quant à elle, suivie de leur comparaison et de l'exposé de pistes proposées dans l'optique d'un prolongement de ce travail.

Peut-on parler d'une « inflation mémorielle » ?

Nous avons voulu savoir si les thématiques mémorielles constituaient, quantitativement, une part importante des associations – autrement dit, il s'agissait de vérifier, de façon sommaire, si l'affirmation concernant une éventuelle « inflation mémorielle » se vérifiait quantitativement.

Premier constat, tel que montré dans les graphiques ci-dessous, le nombre de déclarations contenant des mots-clés relevant du registre mémoriel est largement inférieur à celui du total des déclarations publiées au JO, sur la période 1997-2014. Le plus fort enregistrement de déclarations (tous types d'annonce confondus) en France s'élève à 122 703 déclarations (hors doublons) en 2003 tandis que celui des déclarations à « teneur mémorielle » est de 4 260 déclarations en 2002, soit 4,18% du total. Ce qui est néanmoins non négligeable.
Sur la période concernée, l'ensemble des déclarations connaît une augmentation moyenne de 5,88% et celle des déclarations contenant les mots-clés de 3,64%. Le nombre total de déclarations croît donc davantage que celui des déclarations « mémorielles ».

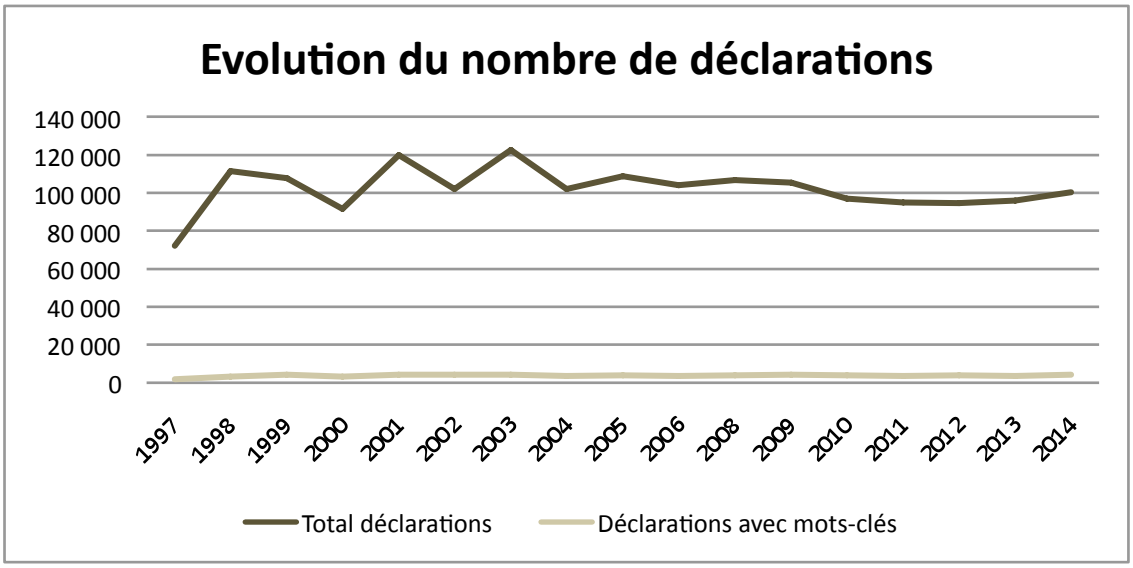


Figure 22 – Nombre de déclarations (1997-2014)

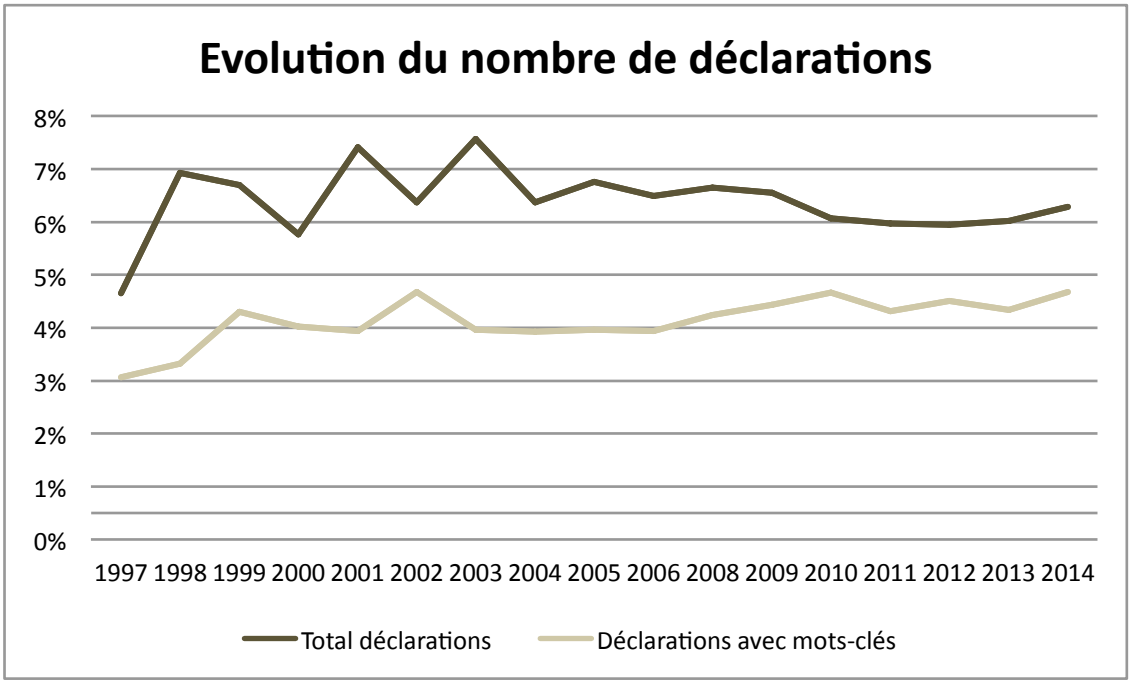


Figure 23 – Nombre de déclarations, en % (1997-2014)

Par ailleurs, l'évolution des déclarations contenant les mots-clés mémoriels fluctue de façon irrégulière. Aucune « inflation mémorielle » massive et linéaire ne semble donc, à première vue, se dégager. Si l'année 2002 représente la plus forte proportion de déclarations avec les mots-clés (4,18%), elle est néanmoins entourée de deux années pour lesquelles le pourcentage des mots-clés est moindre (3,44% pour 2001 et 3,46% pour 2003).
Si nous considérons, maintenant, les annonces contenant le seul mot-clé « Mémoire », la proportion moyenne relativement à l'ensemble s'élève à 0,56% sur l'ensemble de la période, loin derrière les fichiers contenant le mot-clé « Patrimoine » par exemple (2,15%) – qui représentent, rappelons-le, 60% des fichiers filtrés par les mots-clés mémoriels.

Cooccurrence des mots-clés : des pôles thématiques spécifiques ?

Nous avons voulu compléter cette première approche quantitative globale par une étude de la combinaison des mots-clés entre eux[164]. Nous avons vu en deuxième partie que certains mots-clés cooccurraient davantage que d'autres – ce qui avait conduit à l'élimination de « Patrimoine » et de « Célébration ». Nous voulons désormais savoir s'il existe des cooccurrences spécifiques des mots-clés conservés ? Autrement dit, pouvons-nous voir des pôles thématiques émerger ? Et si oui, connaissent-ils des variations dans le temps ?

Les graphiques ci-dessous nous apprennent qu'en 2000, le mot-clé « Mémoire » cooccurre le plus avec « Souvenir » (35% des fichiers), ainsi qu'avec « Anciens combattants » relié à « Souvenir » (19%).
De son côté, le mot-clé « Anciens combattants » apparaît le plus fréquemment en présence de « Souvenir » (54%) et du duo « Mémoire » et « Souvenir » (34%).
Se dessine donc un pôle thématique composé d'anciens combattants, de mémoire et de souvenir. Le terme « Commémoration »ne relève pas, ou plutôt, dans une moindre mesure, de cette sphère mémorielle. A quel autre registre de la mémoire appartient-il ? Avec quels autres éléments lexicaux les mots-clés coooccurrent-ils ?

[164] Cette évaluation quantitative de la cooccurrence des mots-clés concerne la seule année 2000 car elle a été faite avant la sélection des autres années (et le temps nous a manqué pour la poursuivre sur les deux autres années de fichiers traités).

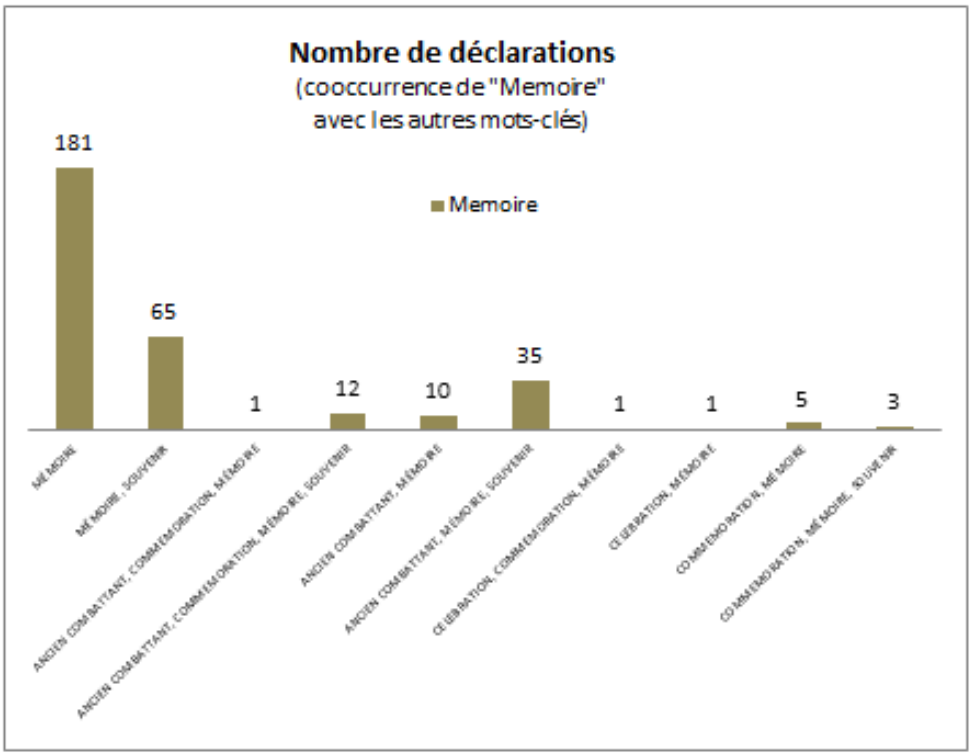


Figure 24 – Cooccurrence de Mémoire avec les autres mots-clés (2000)

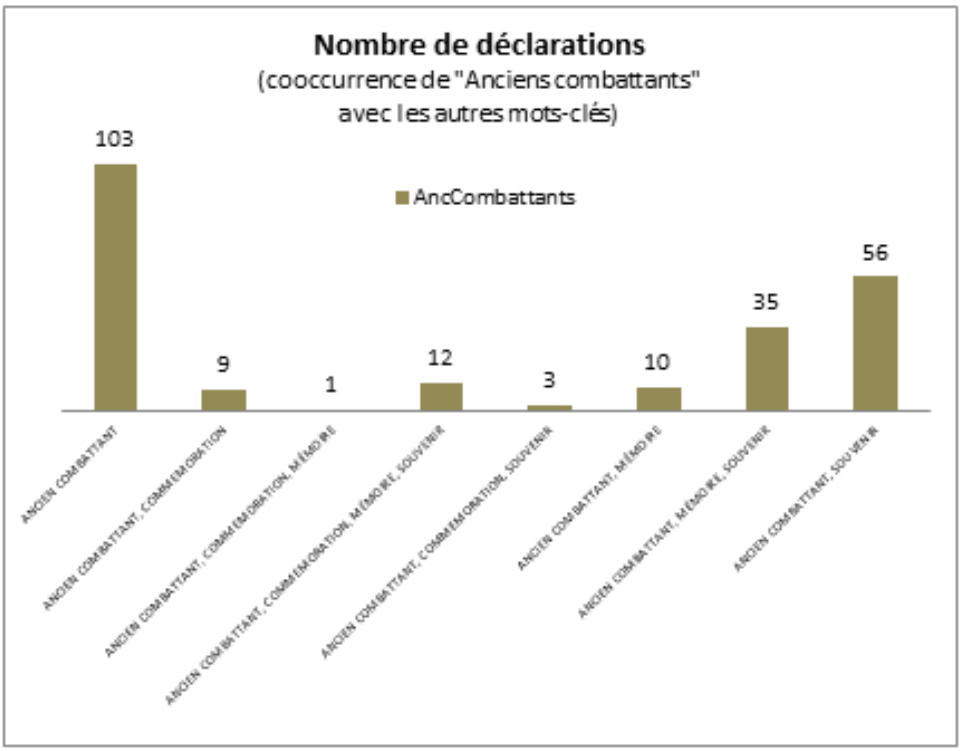


Figure 25 – Cooccurrence de Anciens combattants avec les autres mots-clés (2000)

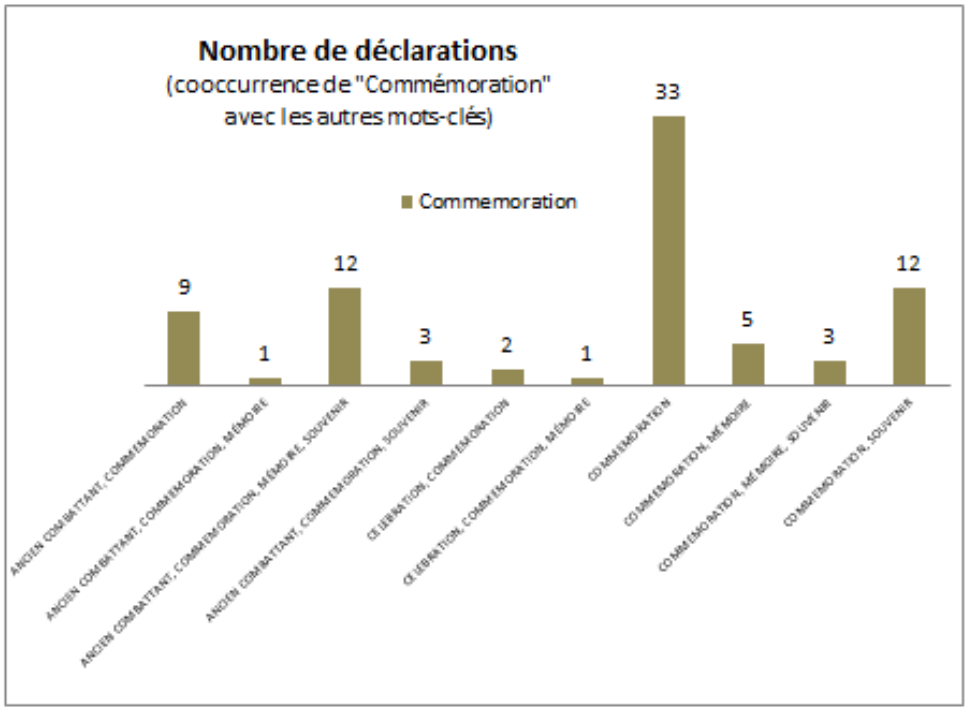


Figure 26 – Cooccurrence de Commémoration avec les autres mots-clés (2000)

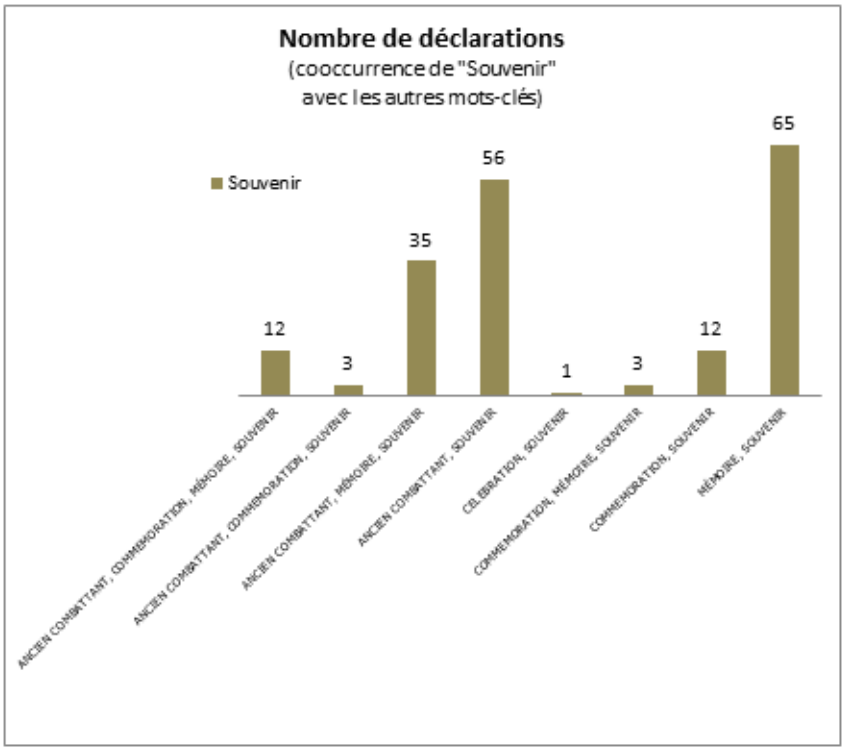


Figure 27 – Cooccurrence de Souvenir avec les autres mots-clés (2000)

Après avoir décrit les relations globales qu'entretenaient les mots-clés entre eux, nous allons détailler ces combinaisons.

Les graphiques[165] ci-dessous présentent l'évolution sur la période des liens de cooccurrence (deux à deux) de « Anciens combattants », « Mémoire », « Souvenir » et de « Commémoration » :

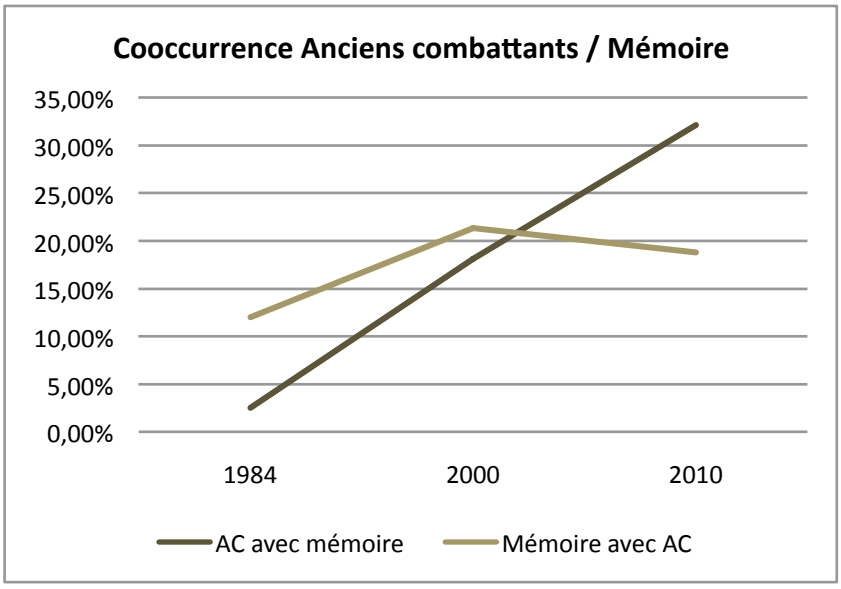


Figure 28 – Nombre de déclaration (cooccurrence AC / Mémoire)

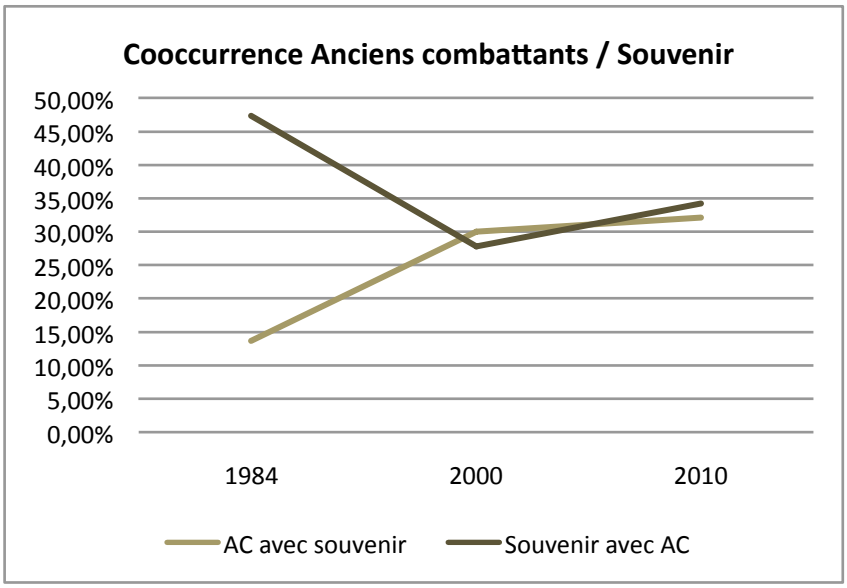
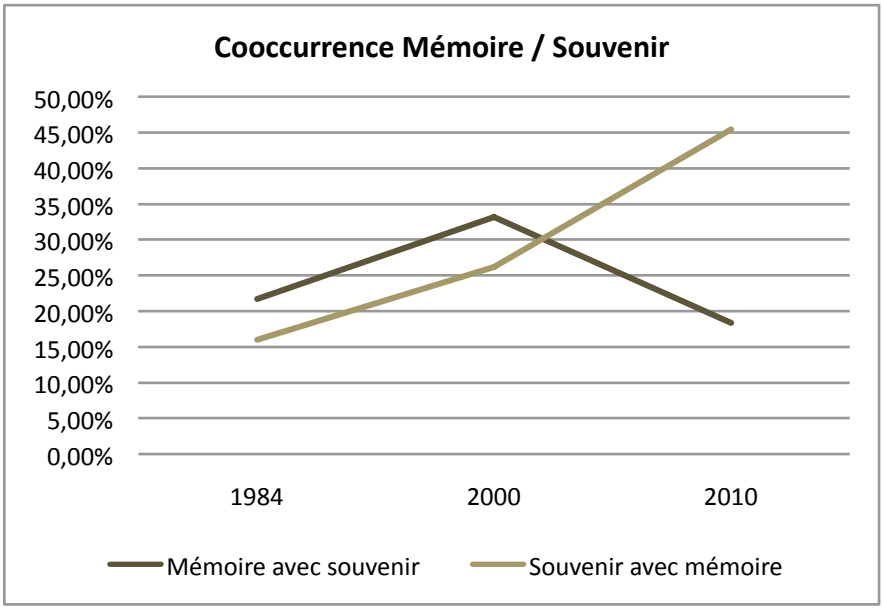


Figure 29 – Nombre de déclaration (cooccurrence AC / Souvenir)

[165] Il faut lire les graphiques de la manière suivante : par exemple, dans le graphique 1, « AC avec mémoire » indique le nombre de fois où le terme AC cooccurre avec celui de mémoire, rapporté au nombre total de déclarations contenant AC (en %) et inversement, « Mémoire avec AC » indique le nombre de fois où le terme mémoire cooccurre avec AC, rapporté au nombre total de déclarations contenant mémoire.



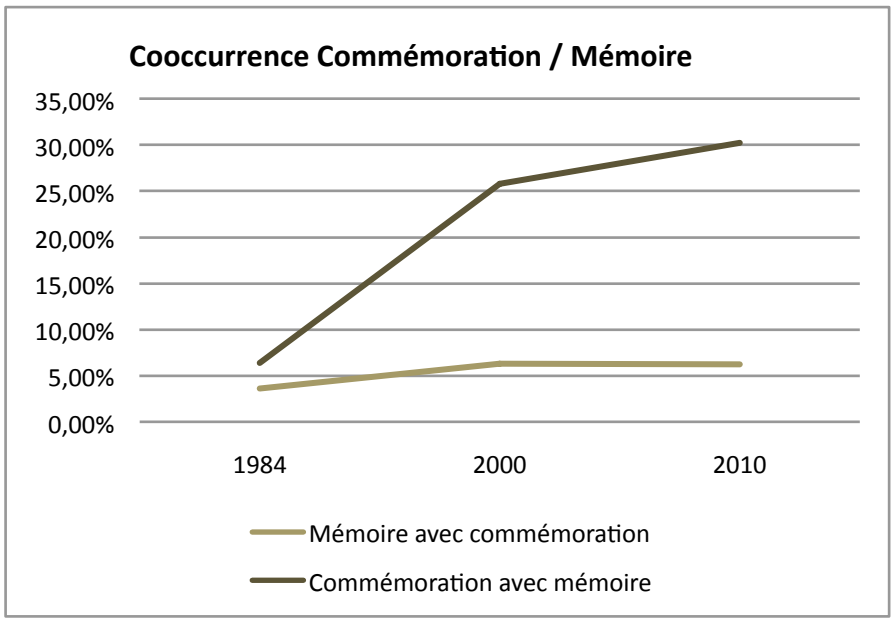


Figure 31 – Nombre de déclaration (cooccurrence Commémoration / Mémoire)

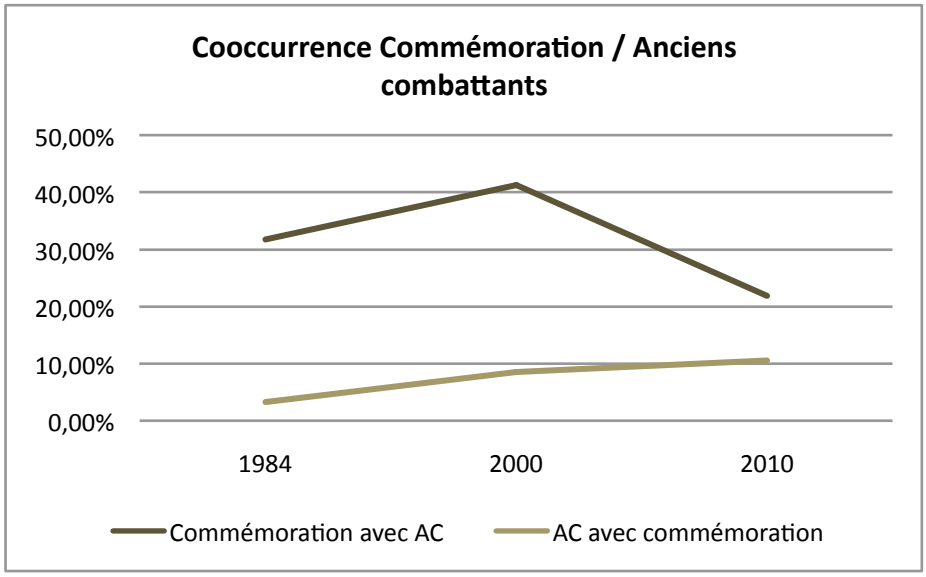


Figure 32 – Nombre de déclaration (cooccurrence Commémoration / AC)

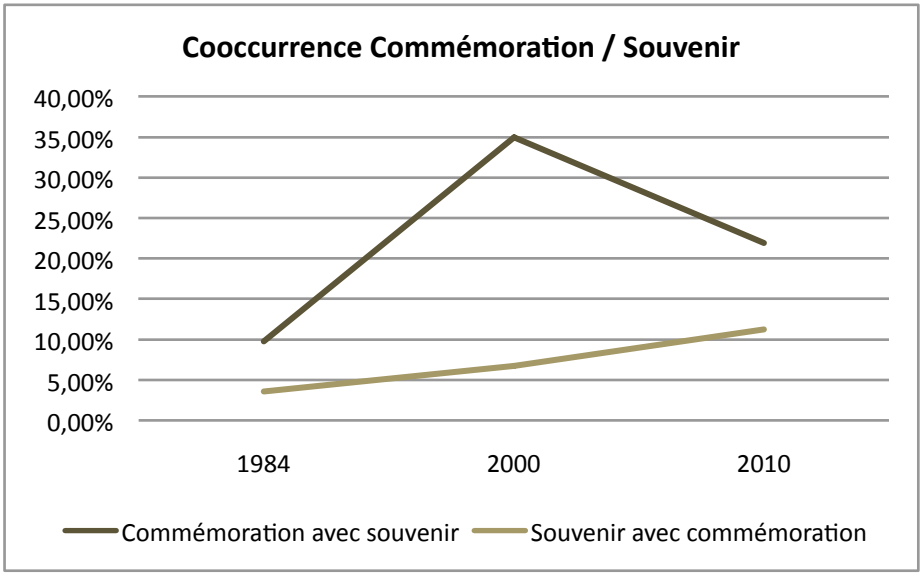


Figure 33 – Nombre de déclaration (cooccurrence Commémoration / Souvenir)

L'analyse temporelle de la cooccurrence des mots-clés entre eux montre des glissements certains entre les termes de la thématique mémorielle. Globalement, en effet, le terme « Mémoire » devient de plus en plus indépendant des autres mots-clés, tandis qu'à l'inverse, ceux-ci tendent à apparaître de plus en plus avec lui. En ce sens, nous pouvons dire que les termes « Anciens combattants », « Souvenir » et « Commémoration » empruntent, dans le temps, davantage au registre de la mémoire que celui-ci ne le fait avec ces termes.
A noter néanmoins que, des 3 mots-clés, « Souvenir » est celui qui cooccurre le plus avec « Mémoire », en début et en fin de période (respectivement, 16% et 45,5%) mais c'est le terme « Anciens combattants » qui connaît la progression la plus forte (+ 30%), passant d'une co-présence quasi inexistante avec « Mémoire » à une très forte cooccurrence. Une substitution a même lieu entre ces deux mots-clés : « Mémoire » prenant la place en 2010 qu'occupait « Anciens combattants » en 1984, à savoir celle du terme le plus « indépendant ». Ainsi, de plutôt isolé en début de période (hormis avec « Souvenir »), « Mémoire » voit ses liens s'accroître sur la période mais dans un sens asymétrique, qui le transforme en mot-clé le plus attracteur.

L'année 2000 constitue un moment charnière dans cette évolution, où la plupart des courbes de cooccurrence des mots-clés s'infléchissent, voire se croisent pour certaines. Ce tournant général comporte néanmoins des différences particulières importantes. En effet, les mots-clés « Anciens combattants » et « Souvenir » voient leur cooccurrence réciproque se rapprocher pour devenir similaire à partir de l'année 2000, alors qu'en 1984, « Souvenir » cooccurrait avec « Anciens combattants » dans la moitié de ses déclarations sans que l'inverse ne soit vrai (« Anciens combattants » apparaissait dans un nombre très important de déclarations[166] sans « Souvenir »). Ainsi, l'indépendance qui caractérisait le mot-clé « Anciens combattants » vis-à-vis du souvenir tend à s'annuler, pour ne plus apparaître que l'un avec l'autre dans un tiers des déclarations les contenant.

Les autres types de corrélations suivent par contre une trajectoire identique : l'un des deux termes de la comparaison cooccurre fortement avec l'autre à partir de 2000, mais sans réciprocité. Ainsi, « Anciens combattants » apparaît conjointement en 2010 avec « Mémoire » dans 32,16% de ses déclarations mais « Mémoire » n'admet « Anciens combattants » que dans 19% des documents le contenant. De même, « Souvenir » cooccurre avec « Mémoire » dans 45,5% des cas en 2010, alors que « Mémoire » n'est présent avec « Souvenir » que dans 18,36% de ses déclarations.

La plus faible corrélation lexicale reliant les 3 mots-clés à « Commémoration », constatée plus haut, se vérifie mais pour l'un des pôles de la relation[167] seulement : quel que soit le mot-clé auquel il est associé, le terme commémoration cooccurre toujours davantage avec lui que l'inverse – même si cet état de fait tend à s'effacer dans le temps, avec une chute brutale à partir de 2000 de ses liens à « Anciens combattants » et « Souvenir ». Tel n'est pas le cas de son lien à « Mémoire », dont l'importance s'accroît fortement (de 6% à 30%), entre 1984 et 2010.

Sur la période considérée, la mémoire s'émancipe[168] donc des autres mots-clés tandis que ceux-ci lui sont davantage inféodés. Reste à savoir quelles modifications précises affectent le vocabulaire mémoriel, outre celui des mots-clés. Seule une analyse détaillée ultérieure du lexique des déclarations et de ses répartitions nous permettra de compléter ce portrait liminaire, voire de le modifier : l'objectif étant, désormais, de vérifier si cette présentation quantitative de l'évolution des mots-clés concorde avec les résultats obtenus à l'aide de Calliope et d'Alceste.

[166] Ainsi, 86,34% des déclarations contenant « Anciens combattants » ne contiennent pas« Souvenir ».
[167] Le taux de cooccurrence de ces 3 mots-clés avec « Commémoration » ne se situe, à son niveau le plus élevé en fin de période, qu'entre 6% et 11%.
[168] Cette émancipation peut également consister en une diversification thématique, doublée d'une augmentation de fréquence du terme au sein du corpus. C'est ce que nous vérifierons plus tard.

2 Analyse des résultats obtenus avec Calliope

	1984	2000	2010
Nombre de documents traités	587	1 107	707
Nombre de descripteurs	299	382	469[169]
Nombre de clusters	7	10	20

Notre présentation des résultats obtenus via Calliope débutera par une description de l'évolution du lexique en général, suivie de l'analyse plus détaillée de la distribution lexicale des 4 mots-clés et de leurs univers thématiques. Ce faisant, nous nous demanderons si des corrélations thématiques se dessinent, si des thèmes sont davantage présents conjointement que d'autres et si ces éventuelles conjonctions varient dans le temps. En un mot, nous tenterons de voir comment la question mémorielle se décline. Par manque de temps, nous n'avons pas pu analyser la distribution géographique des résultats, ni leur analyse en fonction du type d'annonces (création / modification de la déclaration).

2.1 Présentation synthétique

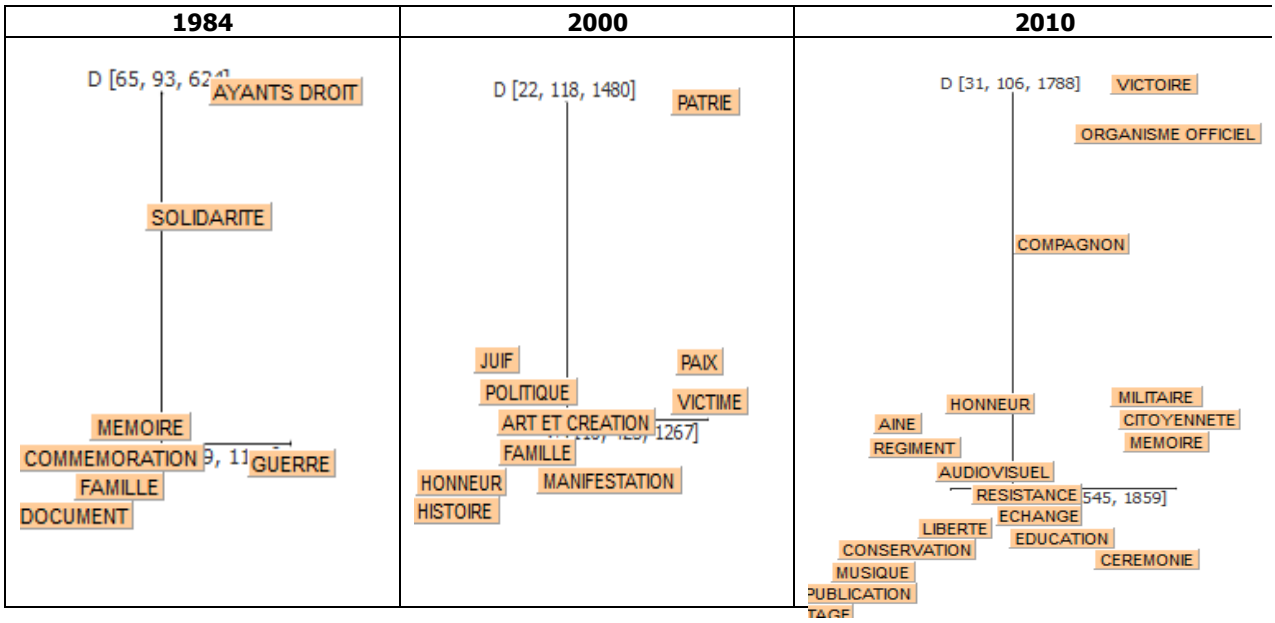


Figure 34 – Diagrammes stratégiques (1984, 2000, 2010)

Une première analyse des diagrammes stratégiques générés par Calliope montre que les thèmes militaires occupent une place importante dans le corpus, sur la période étudiée. Les thèmes représentés par « ayants-droits », « [liens de] solidarité », « patrie », « victimes », « victoire », « organisme officiel », ou encore « militaire » sont tous situés, sur les 3 années,

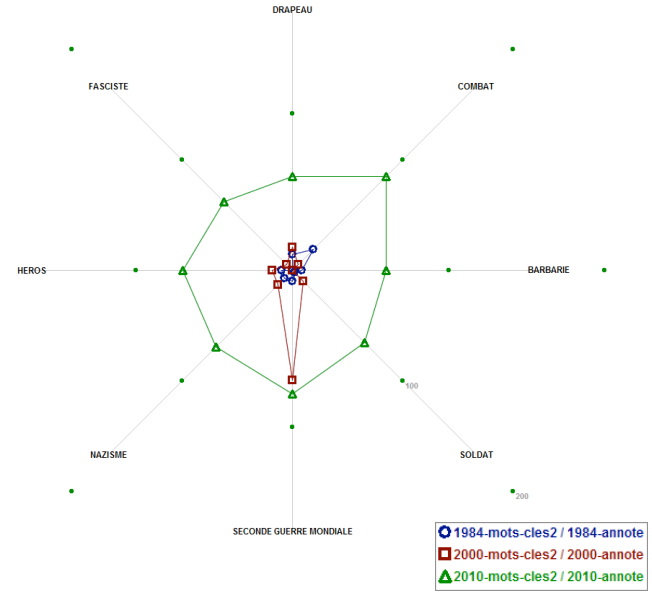
[169] A noter que le nombre de descripteurs en 2010 est supérieur à celui de 2000, pour un nombre de déclarations inférieur. On peut donc en déduire que l'homogénéité des thématiques s'est renforcée.

dans la partie du diagramme qui rassemblent les thèmes phares du corpus, au fort pouvoir structurant et fédérateur. Parallèlement, des thèmes à connotation culturelle, initialement mineurs et peu présents, connaissent un renforcement de leur position au cours de la période. Ainsi, le terme commémoration passe du quadrant sud-ouest (qui regroupe les thèmes émergents) au quadrant nord-est (thèmes stratégiques), relevant du cluster « Militaire ». De même avec mémoire qui devient, en fin de période, l'un des thèmes les plus fédérateurs, formant un réseau lexical très homogène. Les thèmes culturels connaissent un fort développement en 2010, avec un renforcement de la dimension patrimoniale (conservation, musique, publication, audiovisuel). D'autres thématiques, plus symboliques (honneur, liberté, citoyenneté) prennent également de l'ampleur sur la période.
Nous assistons donc à une transformation du lexique mémoriel, qui se traduit par un élargissement de son périmètre lexical (thèmes plus nombreux) et un accroissement de son homogénéité (davantage de liens entre termes). A un ensemble thématique essentiellement caractérisé par l'univers des anciens combattants en 1984, se substitue, au fil du temps, un lexique plus riche, faisant place à d'autres thèmes, notamment culturels et symboliques. Mais il nous faudra analyser plus finement la modification du lien au passé et de sa représentation que cette évolution du vocabulaire suggère.

Cette analyse liminaire peut être complétée par l'étude de la liste des termes qui fluctuent sur la période. Le renouvèlement du lexique des déclarations que nous venons d'indiquer se confirme ici, dans la mesure où davantage de termes émergent (135 descripteurs sur un total de 204) qu'ils ne déclinent (10) sur les 3 années – 43 termes demeurant stables[170]. A noter que les 10 mots qui déclinent sur cette période relèvent quasi exclusivement du domaine militaire, de la guerre (champ d'honneur, défense des droits [des anciens combattants et de leurs ayants-droits], prisonnier de guerre, mutilé et Indochine) mais que d'autres termes en lien à ce même domaine (armée, soldat, combat, drapeau, nazisme, fasciste, pupille, héros, honneur, médaille, mérite, sacrifice, etc.) émergent.

Nous constatons donc une substitution de certaines thématiques militaires à d'autres, plus anciennes. Comment caractériser cette évolution dans la manière de désigner les conflits et leurs participants ? Une première réponse consisterait à indiquer que le lexique récent privilégie, non plus un rapport « direct » aux combattants qui ont été prisonniers de guerre, mutilés ou morts au champ d'honneur, mais une dimension davantage symbolique, insistant sur les actions à entreprendre pour rendre honneur aux combattants qui se sont sacrifiés et reconnaître leur valeur (héros, médaille, mérite). Ces conflits sont situés dans un contexte historique idéologiquement qualifié (fascisme, barbarie).
Reste à vérifier si cette première lecture se vérifie lors de l'analyse détaillée des clusters et d'un retour aux sources.

[170] Il faut entendre par « mots stables » des mots qui apparaissent dans les documents de nos trois sous-corpus (1984, 2000 et 2010) avec un poids supérieur au seuil de spécificité. Si un terme stable appartient à un cluster à chaque sous-période, cela ne signifie pas que son poids ne varie pas sur l'ensemble de la période concernée. Cela signifie uniquement que ce poids est toujours strictement supérieur à ce seuil – contrairement aux termes émergents et aux termes déclinants. Les termes émergents sont donc caractérisés par un gain d'importance sur la période, en ce qu'ils se retrouvent plus fréquemment associés à d'autres termes dans les documents, sans que leur fréquence augmente nécessairement ; les termes déclinants, quant à eux, voient leur poids chuter sur la période. D'un poids suffisant en début de période pour appartenir à un cluster, ils ne participent plus à aucun cluster sur les autres périodes car ils ne cooccurrent plus suffisamment avec d'autres groupes.



Les autres termes émergents relèvent de différents domaines, avec une apparition notable du domaine de l'action sociale (éducation, école, emploi, insertion sociale, reconversion professionnelle, logement, jeunes, personnes âgées, etc.) – qui n'était pas visible à partir des seuls diagrammes –, de celui des droits de la personne (droits de l'homme, égalité, dignité, citoyenneté, humanisme, humanitaire, justice, valeur morale, cohésion, intégration, etc.) ainsi qu'une prise en compte des différents moments de la vie et du lien intergénérationnel (enfance, jeune, parent, personnes âgées, retraites, transmission, avenir, etc.).

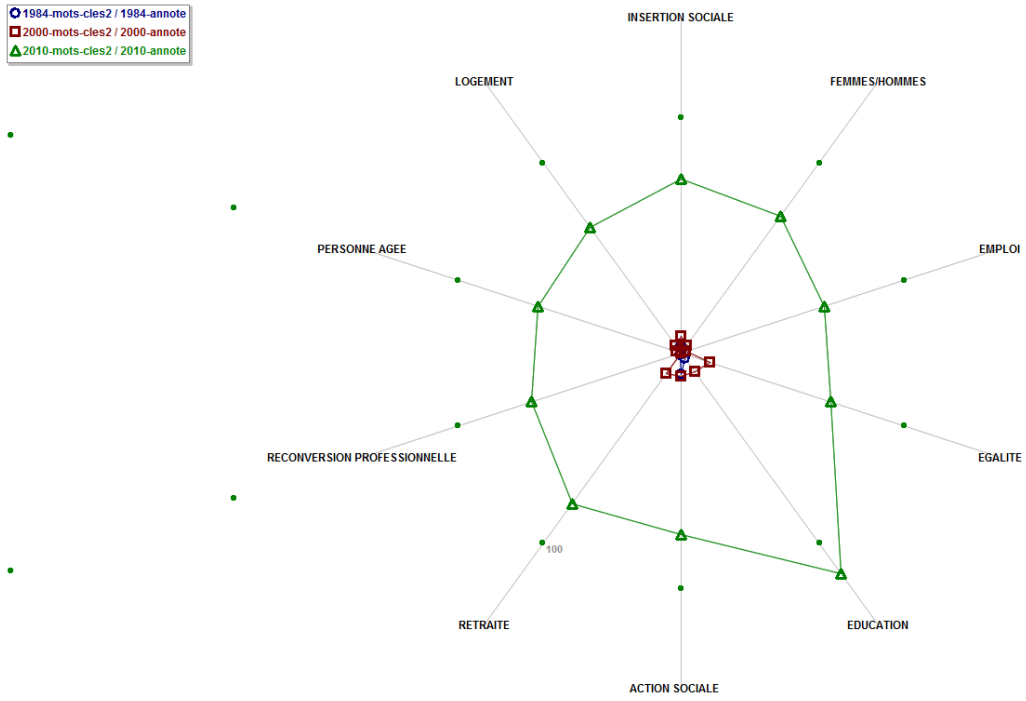


Figure 36 – Mots émergents

Comme constaté plus haut, le thème de la protection patrimoniale, au sens large, fait également une apparition notable (conservation, archive, restauration, tradition,

environnement, développement durable, rural, etc.). Certains termes intéressant directement notre étude se manifestent également (devoir de mémoire, lieu de mémoire, témoignage, antisémitisme, juif, identité, racisme, etc.). Reste à savoir de quelle manière ces termes se rattachent (ou non) aux mots-clés mémoriels.

Sur les 43 mots stables du corpus (ie les termes dont le poids[171] sur la période demeure stable), notons la présence significative des 4 mots-clés mémoriels : « anciens combattants », « mémoire », « commémoration », et « souvenir ». Certains termes en lien avec les anciens combattants sont également stables : « combattants morts », « Afrique du Nord » (qui désignent les anciens combattants d'Afrique du Nord), « guerre », « militaire », etc.

Mais indiquer que les mots-clés demeurent globalement stables sur la période ne suffit pas. Afin de voir si d'éventuels glissements lexicaux affectent la question mémorielle, nous allons désormais analyser les univers thématiques dans lesquels ils apparaissent pour voir ceux-ci connaissent (ou non) des transformations sur la période.

2.2 Analyse de l'évolution des mots-clés mémoriels

Afin d'avoir une vue synthétique de la présence des mots-clés au sein de chaque sous-corpus annuel, nous avons compté le nombre de déclarations les contenant[172] :

	1984	**2000**	**2010**
Anciens combattants	398	469	199
Mémoire	83	398[173]	463
Commémoration	41	97	96
Souvenir	112	504	187

Premier constat, d'un point de vue quantitatif – et ce, malgré la stabilité générale des mots-clés – le lexique des déclarations tend à recourir davantage, sur l'ensemble de la période considérée, au terme « mémoire » et moins à celui d'« anciens combattants ». L'inversion des tendances entre ces deux termes, constatée grossièrement en début de partie, semble se confirmer ici – même si le poids d'anciens combattants demeure important.
Bien que son emploi enregistre une petite hausse sur la période, le terme « Commémoration » a, parmi les mots-clés mémoriels, la fréquence la plus faible ; le vocable « Souvenir » connaît, quant à lui, une fluctuation importante, avec un pic en 2000. Pour pourvoir expliquer cet accroissement temporaire, il faudrait analyser le type

[171] Le poids des termes est une fréquence pondérée, qui fonctionne comme indicateur de la spécificité. Il permet de voir si un terme est surreprésenté ou sous-représenté dans une portion du corpus relativement à l'ensemble du corpus. Pour en savoir plus sur la notion de spécificité, voir LEJEUNE C. et A. BÉNEL, « Lexicométrie pour l'analyse qualitative. Pourquoi et comment résoudre le paradoxe? », *op. cit.*
[172] Précision : ne sont prises en compte dans ce tableau que les déclarations contenant au moins l'un des mots-clés (et non les déclarations ne contenant qu'un mot-clé, à l'exception des autres).
[173] Notons que lorsque nous filtrons les déclarations au moyen du mot-clé « Mémoire » (et ses dérivés « mémoires », « mémoriel » et « mémoriels »), et non plus via l'expression « m ?mo » telle qu'utilisée lors des premiers filtrages sous Excel, le croisement entre les courbes de « Anciens combattants » et de « Mémoire » ne se situe plus en 2000… mais en 2001. En effet, le terme de requête « m ?mo » filtre aussi bien les mots « mémoire » que « commémoration » (et leurs dérivés). En distinguant les deux termes, le nombre de documents indexés varie nécessairement. Cela constitue, comme nous l'avons vu plus, l'une des limites et difficultés de la manipulation de données. Pour des questions de temps, il nous a été malheureusement impossible de traiter le lexique l'année 2001 – ce qui sera fait dans le cadre de la poursuite de ce travail exploratoire.

d'associations concernées et le rapprocher du contexte politico-institutionnel et social dans lequel il s'inscrit.
De façon « évidente[174] », le poids informationnel[175] de 3 mots-clés (« Anciens combattants », « Mémoire » et « Souvenir ») est significatif dans les déclarations dans la mesure où ils figurent, dans les 3 sous-corpus, parmi les 5 premiers descripteurs[176] sur les 204 descripteurs caractérisant le lexique de l'ensemble des fichiers. Sur les 4 mots-clés, un mot-clé (« Mémoire ») fait état d'un positionnement beaucoup plus fort, en termes de valeur informationnelle, en fin de période qu'en début, deux mots-clés demeurent plus ou moins stables (« Commémoration » et « Souvenir », même si ce dernier voit son poids baisser légèrement en 2010) et « Anciens combattants » qui régresse sur la période.

Cette variation temporelle du poids informationnel des mots-clés est représentée dans le graphique ci-dessous :

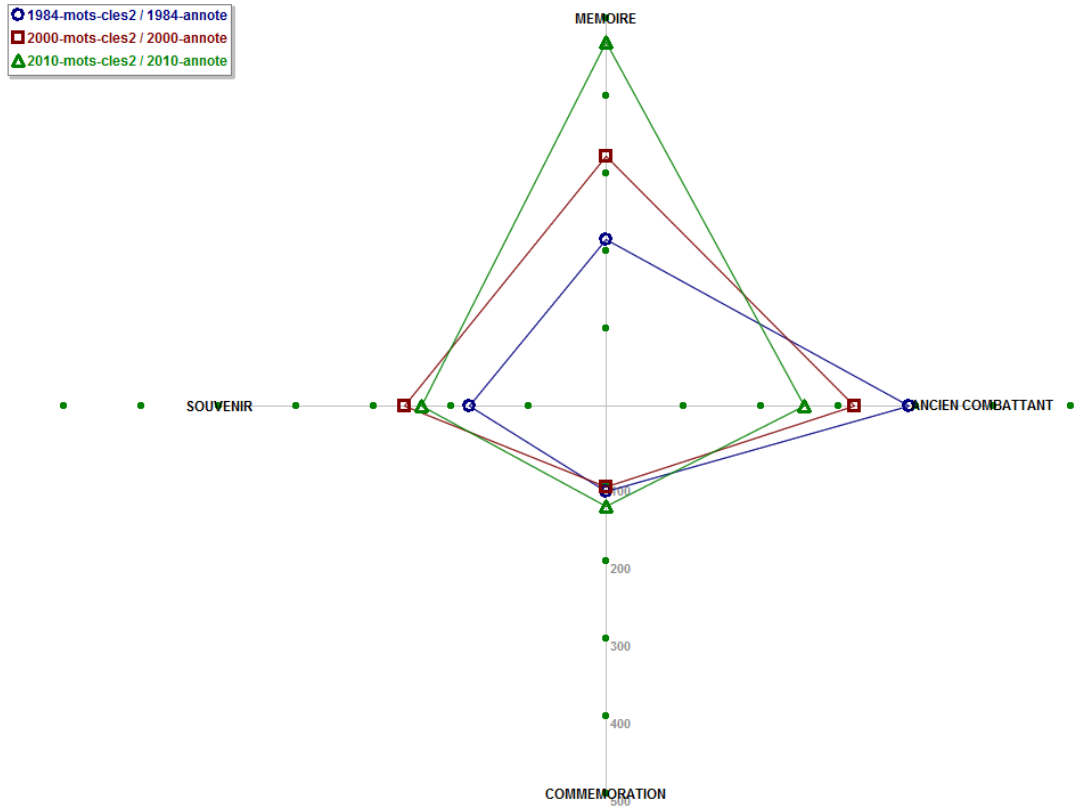


Figure 37 – Evolution des mots-clés mémoriels

Après cette rapide présentation de l'évolution du poids des mots-clés, nous allons désormais détailler les univers thématiques au sein desquels ils apparaissent, ainsi que leurs évolutions.

[174] Dans le cadre d'un travail portant sur le lexique mémoriel.
[175] Le poids informationnel des mots constitue le reflet de leur participation au corpus (calculée comme une fréquence pondérée). Il traduit la participation d'un mot au diagramme stratégique, rapportée à sa fréquence dans l'ensemble du corpus. Par exemple, si deux mots ont une faible fréquence dans l'ensemble du corpus (ie, qu'ils apparaissent peu) mais qu'ils apparaissent systématiquement ensemble (cooccurrence maximale), alors cette particularité augmente leur poids informationnel : la corrélation lexicale corrige ainsi l'importance informationnelle attribuée aux termes que la simple fréquence n'aurait pas réussi à cerner.
[176] Classés selon le poids, de façon décroissante.

2.2.1 Evolution temporelle de « Anciens combattants »

Diagrammes stratégiques et clusters :

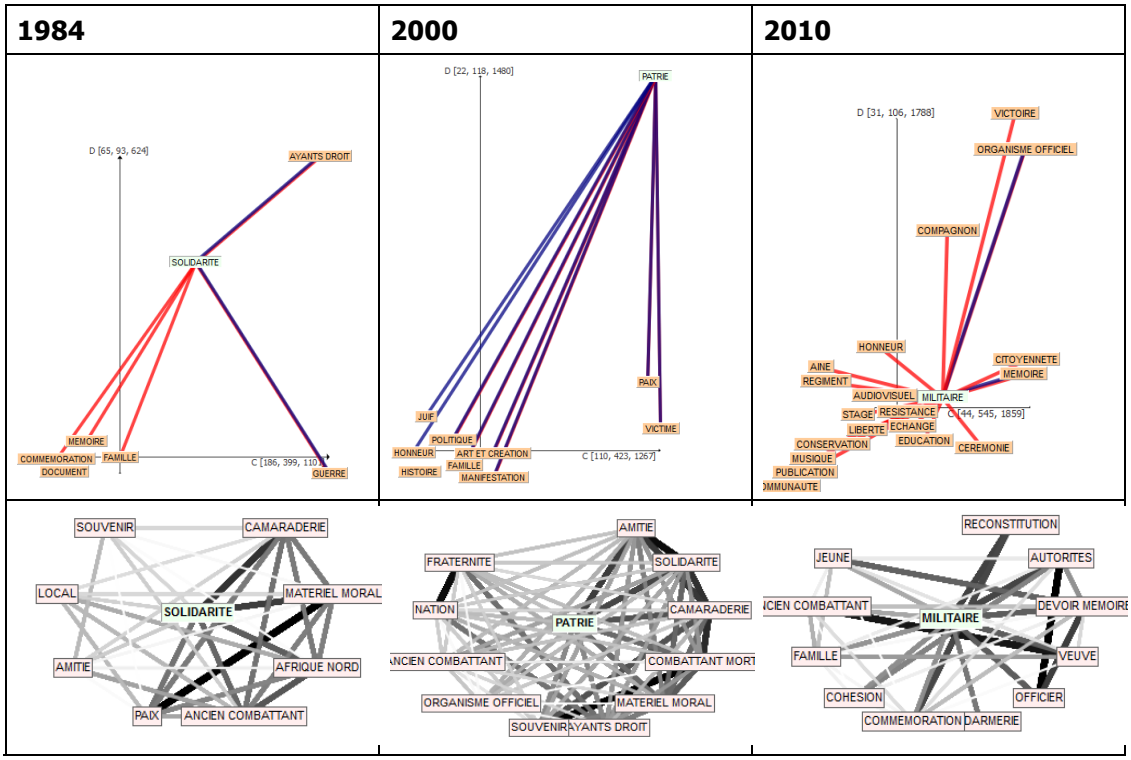


Figure 38 – Diagrammes stratégiques et clusters « Anciens combattants »

Bien que demeurant un terme majeur du corpus, « Anciens combattants » connait néanmoins une baisse de son importance sur la période Ainsi, les clusters auxquels il appartient, tout en restant situés dans la partie du diagramme stratégique qui rassemble les clusters les plus fédérateurs, voient leur couverture du corpus[177] se réduire ainsi que leur densité – malgré un pic notable en 2000. L'année 2000 représente, à ce propos, une année exceptionnelle, avec une densité et une centralité telles que le cluster « Patrie » (duquel relève « Anciens combattants ») en constitue le cluster phare : ses termes sont à la fois les plus présents et les plus liés (connexion intra- et inter-clusters) du sous-corpus. En ce sens, ils possèdent le plus haut degré d'attractivité.

En 1984 et en 2000, l'environnement thématique du terme « Anciens combattants » est très proche, partageant les notions de liens de solidarité, de camaraderie, d'amitié et d'intérêt moral et matériel des ayants-droits. A noter que sa cooccurrence à « Souvenir » sur ces deux années, indiquée plus haut, se confirme.

Un renouvellement lexical caractérise l'année 2010, au sein du cluster « Militaire » : des liens particuliers de cooccurrence se dégagent, reliant notamment anciens combattants et veuves, reconstitution et commémoration, devoir de mémoire et jeune, autorité et officier. Ces thématiques insistent sur les liens familiaux et générationnels (veuve, famille, jeune), un lien qui se traduit en « devoir de mémoire » envers les anciens combattants, qu'il s'agit de commémorer. La commémoration et la reconstitution semblent se substituer ici au souvenir des années précédentes.

En analysant désormais le vocabulaire auquel est relié extérieurement le terme « Anciens combattants », nous nous apercevons que l'on s'éloigne, en quelque sorte, de la sphère thématique typique des anciens combattants (« guerre » et « ayants-droits » représentent

[177] Ainsi, le nombre de documents au sein desquels les mots des clusters dans lequel se situe « Anciens combattants » passe de 51,6% en 1984 à 58,5% en 2000, puis chute à 18,2% en 2010.

87% des liens externes en 1984) pour gagner en épaisseur symbolique et en représentativité : mémoire[178], cérémonie, organisme officiel[179] en 2000 et paix, victime[180] en 2010.
Le lien à « Souvenir », qui a disparu du cluster auquel appartient « Anciens combattants » (« Militaire ») en 2010, est néanmoins maintenu dans la mesure où il apparaît cette année-là au sein cluster « Mémoire », qui représente le lien externe le plus important de « Militaire ». Les deux termes, anciens combattants et souvenir, continuent donc de cooccurrer.

2.2.2 Evolution temporelle de « Mémoire »

Diagrammes stratégiques et clusters :

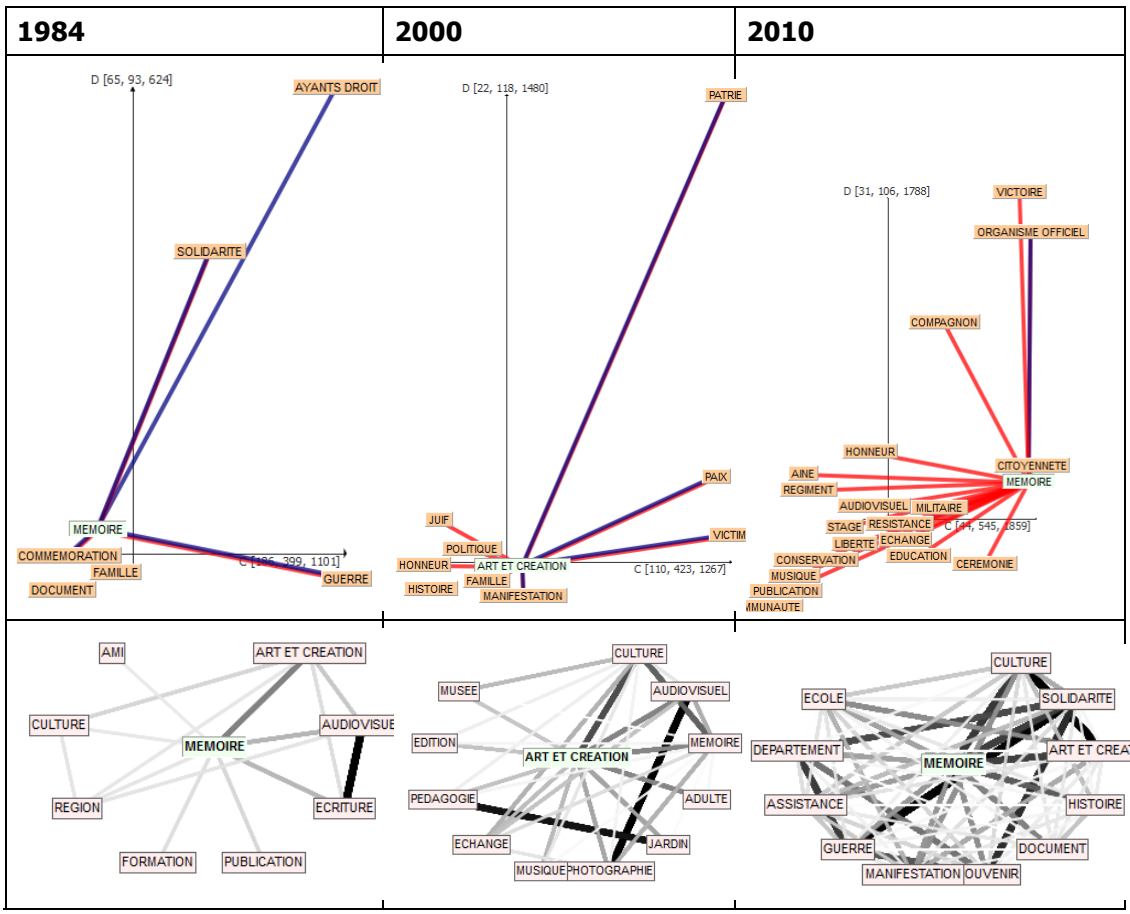


Figure 39 – Diagrammes stratégiques et clusters « Mémoire »

En observant les diagrammes stratégiques et la constitution des clusters des 3 années, nous remarquons que le terme « mémoire » relève principalement du domaine artistique[181], tout

[178] Comme souligné auparavant, le lien de « anciens combattants » au terme « mémoire » s'accroît très fortement sur la période, passant de 0,7% en 1984 à 29% en 2000, puis à 28,8% en 2010.
[179] Ce terme désigne l'Office national des anciens combattants et victimes de guerre (ONACVG), créé en 1916. Ses missions ont évolué au fil des conflits successifs qui ont marqué l'Histoire de la France depuis le début du XXe siècle. Il est intéressant de noter que sa devise actuelle est « "Mémoire et Solidarité" ; elle rappelle que l'établissement public s'investit pleinement dans la préservation des droits matériels et moraux du monde combattant, mais aussi dans la transmission des valeurs de ce dernier. Aujourd'hui, l'ONACVG, est un établissement public, sous tutelle du ministère de la défense. », Site web institutionnel : http://www.onac-vg.fr/fr/onacvg/historique/
[180] Ces deux mots représentent 80% des liens externes.

en se diversifiant avec le temps. D'une thématique émergente en 1984, qui est fermée sur elle-même, peu dense et n'entretient que peu de liens à l'ensemble des documents, elle s'ouvre peu à peu à d'autres univers.
Le cluster « Mémoire » voit ainsi sa cooccurrence avec d'autres termes du corpus s'accroître fortement au fil du temps (de 8,5% en 1984, sa couverture du corpus[182] passe à 62,4% en 2010), de même que sa cohésion thématique – qui connaît à la fois une ouverture à d'autres thématiques et un renforcement de ses liens internes[183]. Cette situation traduit une très forte consolidation de ce cluster dans le temps, dont la progression est l'une des plus importantes du corpus.
Cette tendance est confirmée en 2010, date à laquelle les termes du cluster « Mémoire » ont à la fois sensiblement renforcé leur structuration interne et développé leurs relations aux autres documents (62,4% du corpus sont indexés par ce cluster). Cette évolution se lit clairement en comparant le positionnement du cluster « Mémoire » au sein des trois diagrammes stratégiques : la trajectoire du cluster suit un mouvement vers la droite, passant ainsi du quadrant nord-ouest, qui rassemble les thèmes périphériques, au quadrant nord-est, devenant de ce fait l'un des clusters les plus centraux du corpus. Ce changement de position traduit un accroissement significatif du pouvoir d'attraction de « Mémoire », ie de sa capacité à « agréger d'autres termes pour former un réseau lexical homogène »[184] – qui confirme ce qui avait été avancé plus haut.

Sur l'ensemble de la période, les déclarations employant le terme « mémoire » ont presque toutes un lien au domaine culturel au sens large[185] (que ce soit par la pratique d'activités artistiques ou par l'utilisation de techniques artistiques en vue de valoriser d'autres activités) – dans les 3 clusters, « mémoire » est toujours relié aux termes « art et création » et « culture » – ainsi qu'au domaine de la pédagogie[186]. Largement spécialisé donc à la créativité artistique et à sa promotion[187] en début de période, il voit ses liens avec des thématiques, faiblement présentes ou inexistantes en début de période, se consolider et se diversifier dans le temps[188].
Cette caractérisation culturelle de « mémoire » ne contredit pas des affinités thématiques avec les anciens combattants, dont les liens s'accroissent dans le temps. Nous constatons une augmentation sensible de l'emploi conjoint des deux mots-clés « mémoire » et « anciens combattants » : en 1984, ils apparaissaient conjointement dans seulement 10 documents (sur 587), pour atteindre 85 documents en 2000 (sur 1 107) et 87 documents en 2010 (sur 707). Cette augmentation en valeur absolue correspond également à un accroissement de la part de ces mots-clés relativement à l'ensemble du corpus, passant de 1,7% des documents en 1984, à 7,68% en 2000 et 12,31% en 2010. Cette proximité lexicale est vérifié par l'étude des liens internes et externes de « mémoire ». En 1984 et

---

[181] 59% des documents appartiennent au domaine des activités artistiques proprement dites, ou de la célébration d'artistes morts. A noter qu'un faible nombre de déclarations (5) a pour objet des thématiques en lien avec ce que nous avons désigné les « revendications particularistes » (juif, banlieues, LGBT, femmes).
[182] La couverture du corpus indique la part de l'ensemble des documents du corpus dans lesquels un terme apparaît.
[183] Sur la période, le cluster « Mémoire » gagne à la fois en centralité et en densité ; ce qui explique l'accroissement important de son poids informationnels : il est plus présent et plus lié aux autres termes du corpus.
[184] Mathilde de Saint-Léger, *Ibid*.
[185] Cette dimension culturelle s'expriment à travers la présence des différentes techniques et moyens d'expression artistiques (écriture, audiovisuel, photographie, musique) ainsi que de leurs supports de diffusion (publication, édition, musée).
[186] Les termes « formation », « pédagogie », « échange » et « école » sont présents dans les clusters, indiquant l' idée de transmission que convie le terme « mémoire ».
[187] Notamment via la célébration de la mémoire d'artistes décédés.
[188] La « mémoire » intervient dans des domaines aussi différents que le domaine médical et psychologique (fonction cognitive de la mémoire), informatique (mémoire vive et dure), ethnologique (protection d'un savoir-faire, de techniques), sportif, etc. Selon son contexte d'apparition, le terme mémoire peut adopter différents sens pour désigner différents phénomènes.

2000, il est lié extérieurement pour un tiers de ses liens aux clusters contenant « anciens combattants » (« Solidarité » en 1984 et « Patrie » en 2000). En 2010, le terme « solidarité » apparaît même dans le cluster « Mémoire » : d'externe, cette thématique propre aux anciens combattants (maintenir les liens de solidarité, d'amitié et de fraternité) s'est sensiblement rapproché de la thématique mémorielle, au sens strict.
Ces deux univers cooccurrent donc de plus en plus sur la période considérée (mais avec une dissymétrie dans la relation, comme nous l'avons vu plus haut, dans laquelle « anciens combattants est davantage dépendant de « mémoire » – 30% des liens externes d'anciens combattants pointent, en 2010, vers « mémoire » alors que l'inverse n'est pas vrai dont seulement 7% des liens externes sont en direction de « Militaire »).
Cette évolution traduit une mise au service des techniques artistiques d'enregistrement et de transmission, qui caractérisent le thème « mémoire », à la remémoration de la guerre et de ses acteurs.

D'isolée en début de période, la mémoire devient donc un thème partagé par le lexique de nombreuses déclarations, notamment celles des anciens combattants.

2.2.3 Evolution temporelle de « Commémoration »

Diagrammes stratégiques et clusters :

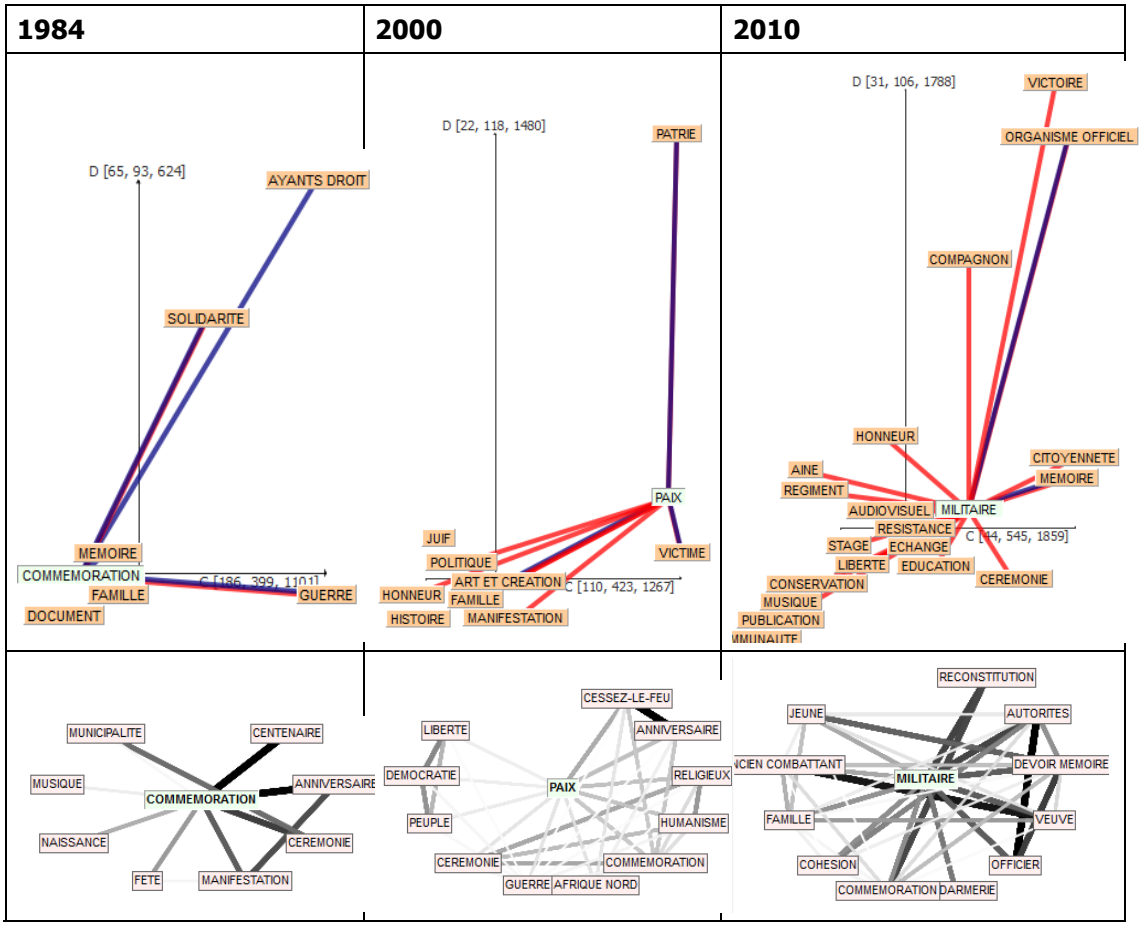


Figure 40 – Diagrammes stratégiques et clusters « Commémoration »

A l'instar du terme « mémoire », le mot « commémoration » connaît une augmentation importante de son poids informationnel sur la période. Le cluster auquel il appartient passe en effet de la partie qui rassemble les thèmes mineurs en émergence (1984) à la zone

contenant les clusters stratégiques du corpus (2000 et 2010). Cette situation un renforcement, traduit des termes auxquels il est associé[189] au sein du corpus et une structuration croissante des thématiques auxquelles il appartient ; autrement dit, les liens de cooccurrence entre les termes du cluster et avec ceux des autres clusters se fortifient – avec, néanmoins, un pic pour l'année 2000. Le terme « commémoration » est ainsi devenu, au fil du temps, de plus en plus attracteur.

En début de période, la commémoration est une cérémonie (manifestation, fête) destinée à célébrer la naissance, l'anniversaire, le centenaire d'un événement, en lien avec la guerre (ce thème représente 50% des relations externes du cluster « Commémoration »), la solidarité et la mémoire. Ces cérémonies peuvent être organisées au niveau des municipalités, pour célébrer des événements d'ordre civil ou militaire.

Cet univers évolue en 2000 pour intégrer dans le cluster « Paix » les valeurs de paix (dont cessez-le-feu), de liberté des peuples, de démocratie, d'humanisme, de patriotisme ainsi que le domaine religieux (cérémonie religieuse). La commémoration se rapporte principalement aux cérémonies d'anniversaire de cessez-le-feu. Cet aspect « culturel » et « événementiel » est confirmé par les liens externes du terme, qui, pour plus de la moitié, concernent le terme « manifestation ». A noter que ces liens renvoient eux-mêmes de façon significative à l'expression « devoir de mémoire » qui émerge cette année-là. La commémoration est donc un événement rappelant, sous forme de cérémonies, des conflits passés (ou plutôt la fin de ces conflits). Cet acte de célébration tend à devenir une obligation à se rappeler, un devoir de mémoire.

Ainsi, en 2010, le terme « commémoration » cooccurre au sein du même cluster (« Militaire ») avec l'expression « devoir de mémoire ». Si l'aspect événementiel est devenu un lien externe (cérémonie, audiovisuel), le rapport aux anciens combattants, aux militaires et à leur mémoire s'est accru. Ce devoir s'exerce en direction des jeunes notamment et la commémoration devient un événement destiné, au moyen de techniques audiovisuelles, à entretenir la mémoire des conflits passés et des acteurs qui y ont pris part. A noter que le mot « reconstitution » relève du même cluster : la cérémonie, plus officielle, a été remplacée par la « reconstitution », peut-être davantage à même de faire revivre les événements historiques et leurs acteurs.

2.2.4 Evolution temporelle de « Souvenir »

Diagrammes stratégiques et clusters :

1984	2000	2010

[189] Les clusters auxquels le vocable « commémoration » appartient indexent davantage de documents sur la période : en 1984, 6,8% des documents contiennent les termes du cluster, 13,2% en 2000 et 18,2% en 2010.

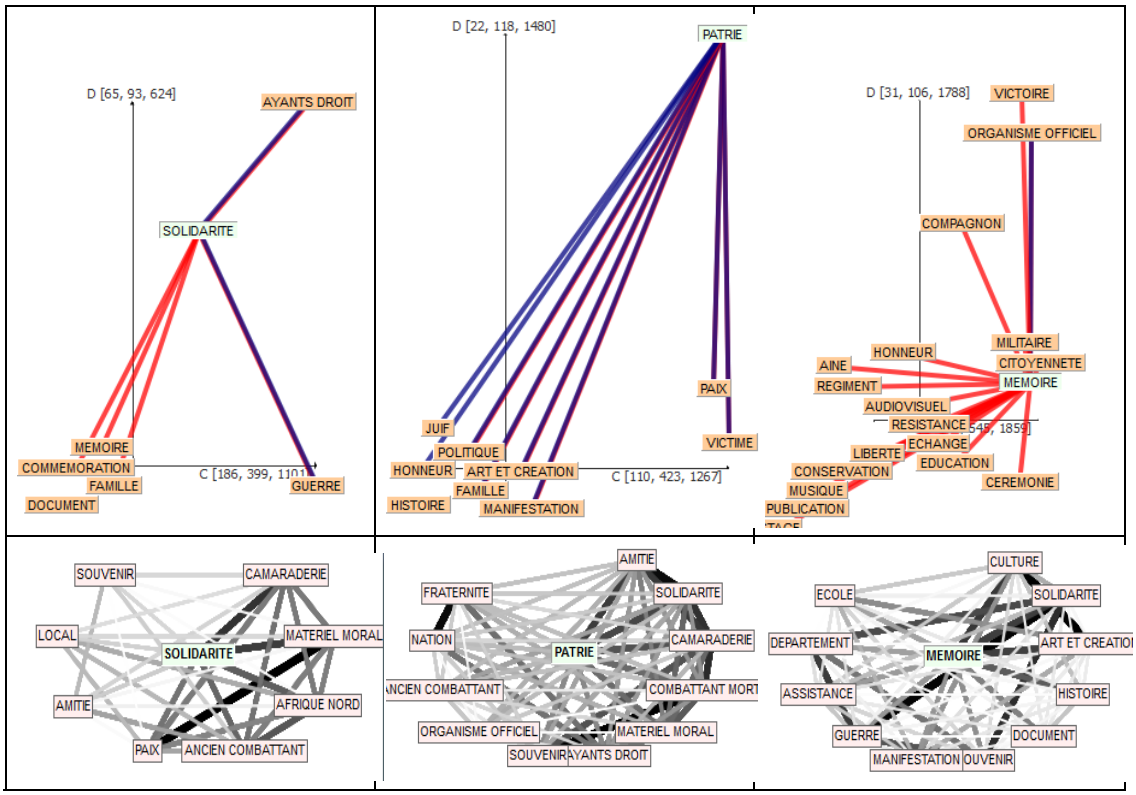


Figure 41 – Diagrammes stratégiques et clusters « Souvenir »

Quelle que soit l'année concernée, le terme de « souvenir » appartient à des thématiques du corpus très présentes[190] et très homogènes : les clusters dont il relève sont toujours positionnés dans le quadrant du diagramme qui réunit les thèmes stratégiques du corpus. En 1984 et 2000, ces thématiques sont composées de notions propres aux anciens combattants (liens de solidarité, d'amitié et de camaraderie, d'intérêt moral et matériel) mais avec un lien à la mémoire, dont elle représente près de 20% de ses liens externes (après « guerre ») en 1984 et près de 40% en 2000.

Ce rapprochement du souvenir à la mémoire se confirme, et l'année 2010 voit même l'intégration du mot souvenir au cluster « Mémoire ». Ce faisant, la dimension proprement mémorielle (et donc culturelle), qui était reliée de façon plus éloignée les années précédentes, devient prédominante en 2010. Nous observons donc un déplacement du « souvenir » de la sphère d'influence des thématiques « typiques » des anciens combattants vers des thématiques toujours en lien avec la guerre (26% des liens externes pointent vers l'ONAC en 2010) mais par la médiation de la mémoire[191]. Si en 1984 il s'agit de perpétuer le souvenir des combattants morts, s'ajoute, en 2010, le fait de servir leur mémoire.

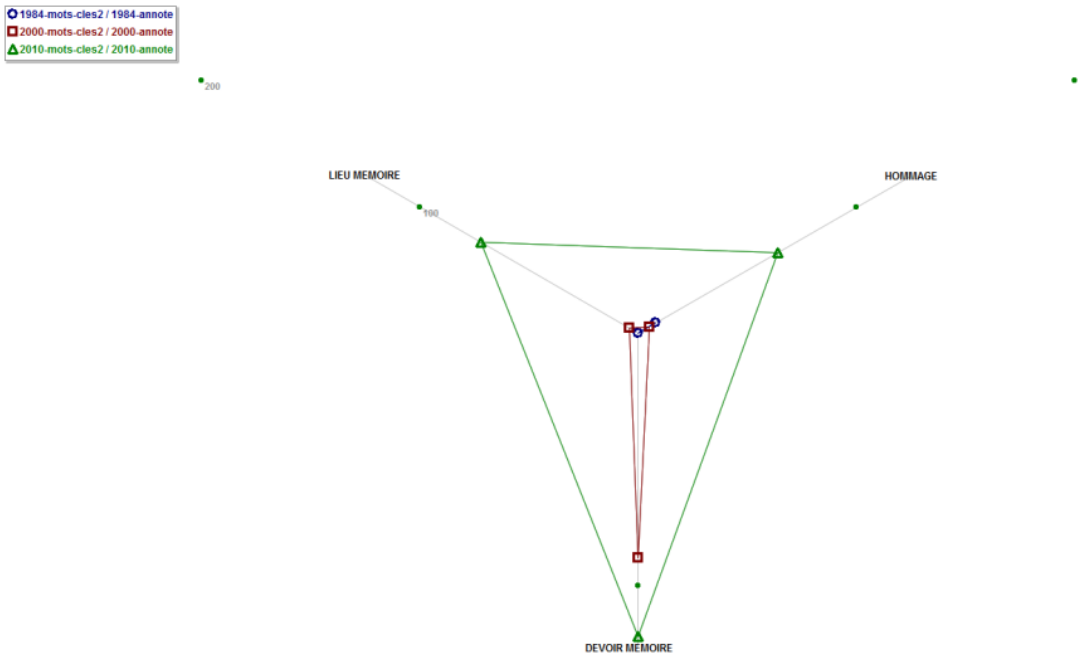
[190] En 1984, le cluster « Solidarité » (auquel appartient « Souvenir ») représente 51,6% des documents ; en 2000, le cluster « Patrie » 58,8% du corpus et en 2010, le cluster « Mémoire », 62,4%. Le terme « Souvenir » lui-même apparaît dans 112 déclarations en 1984, 504 en 2000 et 187 en 2010.

[191] En 1984, ces deux termes apparaissent conjointement dans 18 documents, 132 documents en 2000 et 85 documents en 2010.

2.2.5 Evolution de « devoir de mémoire »

Afin d'affiner cette analyse, nous avons voulu savoir quels étaient les univers thématiques dans lesquels s'inséraient les termes mémoriels, issus des discours institutionnels, tels que « devoir de mémoire », « lieu de mémoire » ou encore « travail de mémoire »[192].

	1984 (587 documents)	**2000** (1 107 documents)	**2010** (707 documents)
Devoir de mémoire / anciens combattants	0 document (0)[193]	32 (17)	63 (19)
Lieu de mémoire / résistance	0 (0)	7 (0)	12 (5)
Travail de mémoire	0	2	2



Tout d'abord, d'un point de vue quantitatif, ces expressions mémorielles ne connaissent pas d'augmentation frappante sur l'ensemble de la période considérée – même si l'expression « devoir de mémoire » connaît la plus forte progression : inexistante en 1984, elle est présente dans 63 documents en 2010 (ce qui représente près de 9% du total). Au sein de notre corpus, leur apparition date de 2000.
Ce « devoir de mémoire » est majoritairement employé, en 2000, dans les déclarations d'anciens combattants : 53% des documents le contenant sont issus d'associations d'anciens combattants, et en représente encore une part non négligeable en 2010 (30% des documents).
Si nous ajoutons au thème des anciens combattants, ceux relevant de manière plus ou moins large du domaine du conflit et de ses acteurs, nous obtenons, pour 2000, une proportion de plus de 80% des documents qui font appel au « devoir de mémoire » (sous-officiers en retraite, visant à préserver le lien entre l'armée et la Nation, résistance durant la seconde

[192] Notons que les expressions « question mémorielle » et « loi mémorielle » sont absentes des descripteurs du corpus – peut-être en raison de leur dimension politico-institutionnelle, trop éloignée du périmètre d'activité des associations.
[193] Les chiffres entre parenthèses indiquent le nombre de déclarations dans lesquelles se trouvent conjointement les deux expressions (comme, sur la première ligne, « devoir de mémoire » et « anciens combattants »).

guerre mondiale, mémoire des républicains espagnols, crimes contre l'humanité commis au Congo, etc.).
Le devoir de mémoire s'avère donc une expression essentiellement liée à la mémoire des conflits passés et à la valorisation des combattants et victimes, qu'ils soient militaires ou civils.
De même en 2010 – mais dans une moindre proportion –, le devoir de mémoire est associé dans un tiers des déclarations aux anciens combattants. En lui adjoignant les documents relevant de la remémoration des conflits passés et de la célébration de leurs acteurs, cette proportion atteint 65%. Si l'usage de l'expression « devoir de mémoire » est relativement moins fréquent en 2010 au sein des associations officielles d'anciens combattants français, il se diffuse néanmoins au sein d'associations orientées majoritairement vers l'univers des combats (mémoire de différents corps d'armée, célébrations des héros américains de la seconde guerre mondiale, des orphelins de guerre, du mur de l'Atlantique, de la résistance, de l'histoire militaire ou encore du Rwanda, des républicains espagnols et des harkis).
A noter que des déclarations prennent – dans une proportion faible, il est vrai – pour objet le devoir de mémoire en tant que tel ou de manière très lâche et imprécise. C'est le cas d'une association de financement de voyages scolaires dont l'intitulé est « Devoir de mémoire », sans qu'il soit précisé dans l'objet le lien entre le titre de l'association et son objet ou, d'une association sur la pêche professionnelle dont le but est « la défense, la promotion de la pêche professionnelle, le devoir de mémoire », ou enfin de l'association « Gardons la mémoire » qui entent « organiser des événements tels que expositions, conférences ou toutes actions mettant en valeur notre devoir de mémoire envers les générations futures ». Le devoir de mémoire semble ici se transformer en stéréotype langagier, dont la simple présence vaut sens.
De son côté, si l'expression « lieu de mémoire » est employé de façon plus diversifiée[194] que celle de « devoir de mémoire », elle entretient néanmoins un lien fort, dans les déclarations de 2010, avec la résistance, et de manière plus générale, avec la seconde guerre mondiale (50% des documents relèvent de ces thématiques).
Les deux déclarations contenant le « travail de mémoire » traitent, pour l'une, des anciens combattants et des victimes du nazisme et pour l'autre, de l'histoire d'un territoire (urbain et rural).

2.2.6 Thématiques « communautaristes »

Dans l'optique du projet de recherche dans lequel nous nous inscrivons, nous nous sommes demandé si nous assistions, tel que l'affirme la croyance générale, à une inflation des « revendications particularistes » ? Pour cela, nous avons étudié, sur l'ensemble de la période à notre disposition, la cooccurrence du terme mémoire avec différents termes, présents dans le corpus, relevant du registre des « identités communautaristes ».

Voici les résultats détaillés, qui indiquent le nombre de documents dans lesquels apparaissent conjointement le terme « Mémoire » et l'un des termes issus de la thématique « communautariste » :

	1984 (587 documents)	**2000** (1 107 documents)	**2010** (707 documents)
Mémoire / racisme	0 document (0)[195]	3 (4)	10 (10)
Mémoire / antisémitisme	0 (0)	1 (1)	4 (4)
Mémoire / juif	2 (2)	5 (5)	5 (6)
Mémoire / Shoah	0 (0)	0 (0)	2 (2)

[194] Elle concerne, entre autre, la préservation de lieux en lien à la mémoire ouvrière, à la biodiversité, à des événements historiques, à des personnalités, etc.
[195] Les chiffres en bleu et entre parenthèses indiquent le nombre total de déclarations contenant au moins un terme « communautariste ».

Mémoire / musulman	0 (3)	0 (0)	1 (3)
Mémoire / immigration	0 (0)	3 (4)	2 (2)
Mémoire / migration	0 (0)	0 (0)	1 (1)
Mémoire / émigration	0 (0)	0 (0)	1 (1)
Mémoire / banlieue	1 (1)	0 (1)	0 (0)
Mémoire / Bretagne	1 (2)	0 (1)	1 (2)
Mémoire / breton	0 (0)	1 (1)	0 (0)
Mémoire / Corse	1 (6)	0 (2)	1 (1)
Mémoire / DOM TOM	0 (0)	17[196] (25)	18[197] (21)
Mémoire / esclave	0 (0)	0 (0)	1 (1)
Mémoire / identité	1 (1)	4 (4)	12[198] (15)
Mémoire / LGBT	0 (0)	0 (0)	3 (4)

D'après cette analyse – et bien que les documents contenant ces thèmes soient extrêmement peu nombreux au regard du nombre total de déclarations –, si aucune inflation « communautariste » ne se dégage[199], il n'en demeure pas moins que la cooccurrence entre le terme « mémoire » et ces autres termes est très forte. Cette corrélation lexicale varie en effet de 100% de proximité (les deux termes apparaissent toujours conjointement au sein des déclarations, comme c'est le cas notamment en 2010 avec mémoire et racisme, mémoire et antisémitisme ou mémoire et Shoah) à 50% à son plus faible niveau, en passant par des scores élevés (83% pour mémoire et juif en 2010, 75% pour mémoire et racisme en 2000). Soulignons que ces chiffres, élevés en proportion, sont néanmoins extrêmement faibles rapportés au nombre total de documents constituant les sous-corpus. Ainsi, les 10 documents comprenant mémoire et racisme en 2010 ne représentent qu'1,47% du total. Et le taux le plus élevé concerne des déclarations contenant le vocable DOM TOM[200] en lien, non pas avec la mémoire de l'esclavage, mais avec celle des anciens combattants.
Si nous rappelons que les termes « mémoire » et « anciens combattants » apparaissaient conjointement au sein de 10 documents en 1984, 85 documents en 2000 et 87 documents en 2010, alors la thématique mémorielle semble être davantage l'apanage des anciens combattants que des associations « communautaristes ».

[196] A noter que, sur ces 17 documents, 15 documents concernent des associations d'anciens combattants. Aucune référence à l'esclavage n'est faite dans ces déclarations (2000 et 2010).
[197] De même, sur ces 18 documents, 15 documents relèvent également des anciens combattants.
[198] Ce thème est extrêmement diversifié puisqu'il contient des déclarations d'associations d'anciens combattants et d'autres portants sur les personnes âgées, la culture berbère, la guerre d'Espagne, les habitants de Paris, les personnes LGBT, les droits de l'homme, etc.
[199] En termes de fréquence ; le poids informationnel de ces mots s'accroît néanmoins dans le temps.
[200] Il faudrait davantage parler de « Groupe maître », selon la terminologie de Calliope, que de vocable car le terme « DOM TOM » regroupe les termes suivants du lexique : Guadeloupe, Guadeloupéenne, Guyanais, Martinique et Martiniquaise, Antillais, Réunionnais et Outre-Mer.

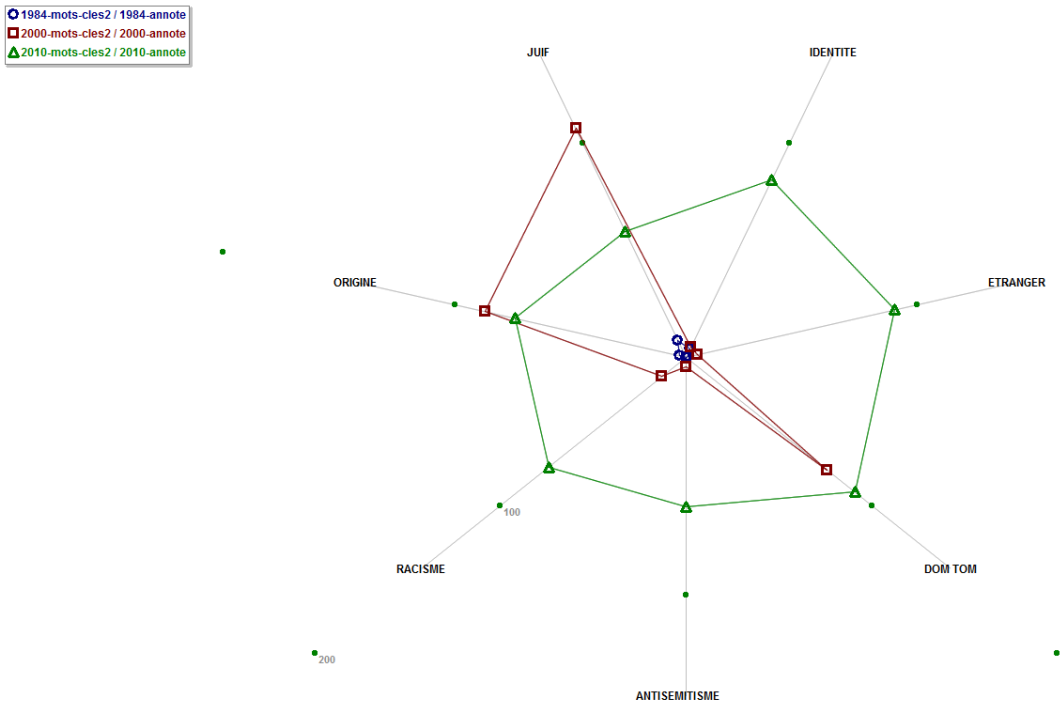


Figure 43 – Evolution des thématiques "communautaristes"

Nous avons également vérifié le lien que pouvait entretenir ces « communautés » à la commémoration. En analysant le tableau ci-dessous, il apparaît que ce terme mémoriel est encore moins employé par ces associations que celui de mémoire :

	1984 (587 documents)	**2000** (1 107 documents)	**2010** (707 documents)
Commémoration / racisme	0 document (0)	0 (4)	0 (10)
Commémoration / antisémitisme	0 (0)	0 (1)	0 (4)
Commémoration / juif	0 (2)	1 (5)	1 (6)
Commémoration / Shoah	0 (0)	0 (0)	0 (2)
Commémoration / musulman	0 (3)	0 (0)	0 (3)
Commémoration / immigration	0 (0)	0 (4)	0 (2)
Commémoration / migration	0 (0)	0 (0)	0 (1)
Commémoration / émigration	0 (0)	0 (0)	0 (1)
Commémoration / banlieue	0 (1)	0 (1)	0 (0)
Commémoration / Bretagne	0 (2)	0 (1)	0 (2)
Commémoration / breton	0 (0)	0 (1)	0 (0)
Commémoration / Corse	1 (6)	0 (2)	0 (1)
Commémoration / DOM TOM*	0 (0)	0 (25)	2 (21)
Commémoration / esclave	0 (0)	0 (0)	0 (1)
Commémoration / identité**	0 (1)	0 (4)	3 (15)
Commémoration / LGBT	0 (1)	0 (0)	1 (4)

Les deux documents associant « commémoration » et DOM TOM traitent d'amicales en lien avec l'armée (amicale de parachutistes en Martinique et amicale régimentaire à la Réunion).

Les déclarations utilisant conjointement le terme « commémoration » et un terme « communautariste » se déclinent comme suit : trois déclarations en lien avec le terme « identité » – il s'agit de commémorer les événements marquants des communautés togolaise, mongole et LGBT (cette dernière faisant, notamment, référence aux déportations de personnes homosexuelles durant la seconde guerre mondiale) ; et deux déclarations utilisant le terme « juif » (pour commémorer le souvenir d'enfants morts en déportation et pour favoriser la diffusion du judaïsme).

Par manque de temps, nous n'avons malheureusement pas pu procéder à l'analyse de la distribution géographique (départements) des résultats ou par type d'annonce (création / modification d'une déclaration).

3 Analyse des résultats obtenus avec Alceste

Notre démarche, dans l'utilisation du logiciel Alceste, se décline en deux étapes.
Nous avons voulu, tout d'abord, analyser conjointement les 3 années de déclarations d'associations, en les réunissant au sein d'un corpus unique, de manière à dégager les distinctions qui caractérisaient, dans leur ensemble, les univers de discours propres à la question mémorielle. Une analyse de la répartition de ces discours quant aux variables date, département et types d'annonce sera également présentée.
Puis, dans une perspective plus dynamique, nous avons procédé à une analyse des univers lexicaux de chaque année individuellement de manière à, non seulement, pouvoir comparer les résultats obtenus avec ceux de la première étape et voir si des recoupements ou des divergences apparaissaient mais surtout, questionner l'influence de la dimension chronologique sur la constitution des univers de discours.
Avant de présenter les résultats obtenus via le logiciel Alceste, nous aimerions insister sur une précaution méthodologique. Le choix de regrouper, dans un premier temps, les 3 années de déclarations en un unique corpus global n'est pas anodin dans la mesure où cela influence la signification accordée à la variable « temps ». Dans cette démarche, l'évolution des discours dans le temps est en effet appréhendée comme une variable illustrative, supplémentaire : cela revient à « considérer implicitement le changement comme second, à l'apprécier par rapport à une structure [83, LEMERCIER et ZALC] ». Faire ce choix présuppose donc que les fichiers des 3 années seraient davantage caractérisés par les divers univers de discours qui les composent que par leur évolution temporelle – ou, du moins, que cette évolution temporelle serait secondaire par rapport à une structuration interne des discours. C'est pour cette raison que nous avons complété cette première analyse par une seconde, afin de voir si la dimension chronologique modifiait les résultats. Est-ce que les spécificités discursives repérées lors de la phase 1 sont affectées lorsque nous introduisons une dimension temporelle ?

3.1 Corpus unique composé des trois années

Considéré globalement, le corpus des déclarations d'associations au JO fait état de deux univers de discours distincts[201] de la question mémorielle dont les énoncés s'opposent, en raison de la cooccurrence de leur vocabulaire (certains termes apparaissent plus fréquemment ensemble qu'avec d'autres termes, ce qui concoure à la distinction des classes).
Ces deux univers constituent deux postures différentes quant au partage et à la transmission de valeurs et aux liens qui réunissent les membres d'une communauté. Un premier domaine de la classification (présentée ci-dessous), au périmètre clairement défini, rassemble les acteurs des conflits passés et les multiples liens qui les unissent, comme les liens de solidarité et de camaraderie (classes 1 et 2) ; l'autre branche, beaucoup plus disparate, qui relève notamment du domaine culturel et symbolique, regroupe les acteurs autour des notions d'hommage à rendre à un certain passé, douloureux (classe 3), de conservation et de valorisation patrimoniale du passé (classe 4) et de l'échange et la transmission de pratiques (culturelles, artistiques, sportives, etc.) entre membres d'une communauté (classe 5).
A noter que ces différentes classes reprennent la catégorisation des mots-clés mémoriels : la classe 1 correspond au mot-clé « Anciens combattants », la classe 2 à « Souvenir », la classe 3 à « Commémoration » et la classe 4 à « Mémoire ». Si cette répartition confirme à postériori la spécificité des discours attachés aux mots-clés (chaque mot-clé renvoie à un univers thématique propre), elle révèle peut-être également la non pertinence de filtrer les déclarations à l'aide de

[201] Rappelons que la méthode implémentée au sein d'Alceste a pour finalité de repérer, statistiquement, des spécificités discursives, ie des proximités lexicales propres à un univers de discours. La constitution de classes repose donc sur l'analyse de la cooccurrence des termes au sein de fragments.

ces mots-clés. Que nous apprend en effet l'analyse des discours liés à ces mots-clés ? Que leurs discours relèvent bien d'univers distincts. Nous retrouvons donc dans les résultats, de manière plus fine, ce qui a été initialement filtré. Cette confirmation, pour utile qu'elle soit, n'en demeure pas moins circulaire. Nous verrons, à l'issue de ce mémoire, les conséquences qui ont été tirées de ce résultat et les choix élaborées en vue de sortir de cette circularité.

Classification[202] – corpus global :

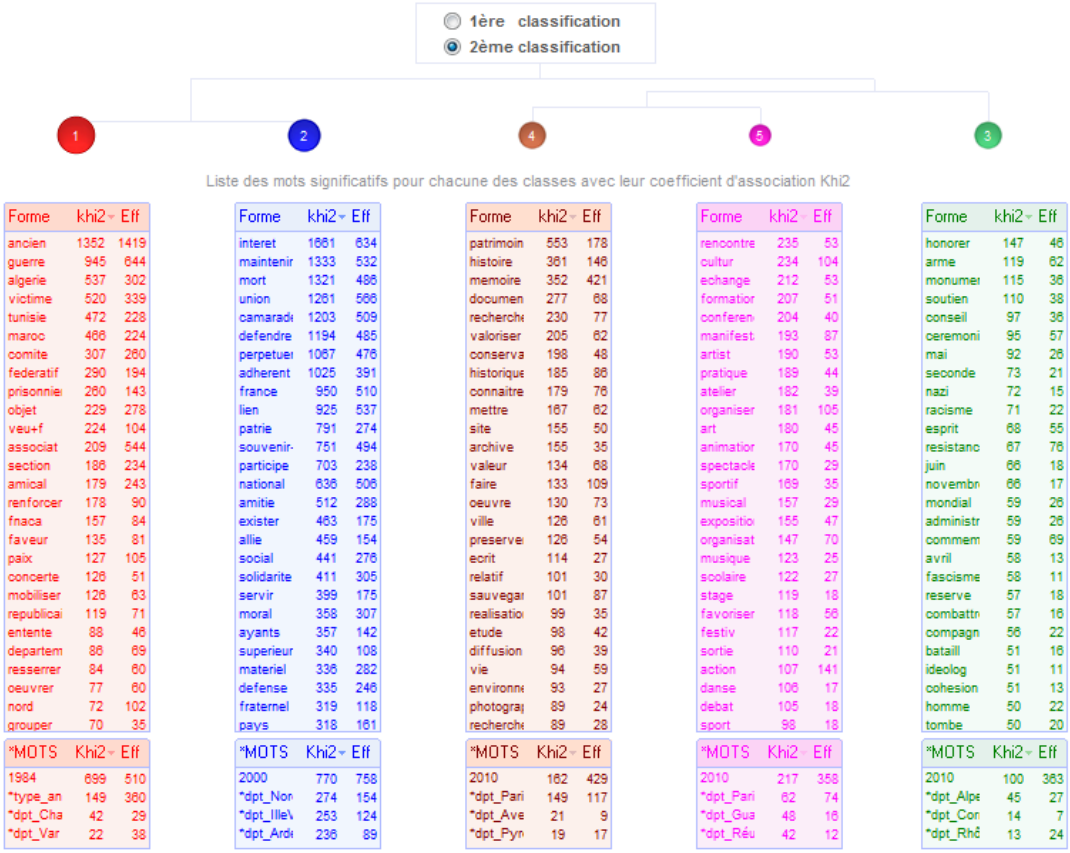


Figure 44 – Classification descendante hiérarchique du corpus global

Informations quantitatives sur le corpus traité par Alceste :

Pourcentage du corpus analysé[203]	83%
Pourcentage classe 1 « Anciens combattants »[204]	30%
Pourcentage classe 2 « Souvenir »	23%
Pourcentage classe 3 « Commémoration »	16%
Pourcentage classe 4 « Mémoire »	17%

[202] Les classes réunissent les « formes qui leur sont le plus spécifique, classées par ordre de significativité décroissant (Lebart et Salem, 1994) », *in* GARNIER B. et F. GUÉRIN-PACE, *Appliquer les méthodes de la statistique textuelle*, *op. cit.*

[203] La part du corpus analysé correspond au pourcentage d'unités de contexte traitées et classées. Lorsque ce taux est supérieur à 70%, il est considéré comme étant pertinent. Autrement dit, le nombre de formes distinctes par UC (sur le nombre total de mots formant les UC) retenu pour le calcul des classes représente bien la diversité (taille et richesse du vocabulaire) du corpus. Si tel n'était pas le cas, le pourcentage du corpus analysé serait faible (le nombre de mots différents retenus par UC serait largement inférieur au nombre total de mots des UC).

[204] Pour plus de clarté, les classes sont ici désignées par le mot-clé mémoriel qu'elles contiennent.

Pourcentage classe 5 « Echanges de pratiques »	14%
Nombre total d'occurrences	142 957
Nombre total de formes distinctes	10 243
Nombre de hapax[205]	5 313
Richesse du vocabulaire[206]	99,65%

Nous remarquons que la première branche de l'arbre hiérarchique (qui rassemble les classes « Anciens combattants » et « Souvenir ») représentent non seulement la part majoritaire du corpus traité (53% de couverture du corpus) mais aussi les classes les plus spécifiques, ie les plus dissemblables aux autres classes. La spécificité maximale qui caractérise cette branche s'explique par la méthode employée par Alceste : la classification descendante hiérarchique. Ce type de classification constitue une méthode de discrimination, qui procède par découpage du corpus selon le degré de spécificité interne et de dissemblance externes des fragments de texte. Plus des fragments seront spécifiques par rapport aux autres fragments, et plus leur constitution en classe sera avancée dans le processus de classification. Avant même de procéder à l'analyse de leurs éléments linguistiques, nous pouvons donc déjà en déduire que les premières classes constituées sont plus spécifiques que les dernières classes. Cette information est utile pour relativiser la cohérence de l'univers de sens décrit par les classes, notamment les dernières.

Deux pôles mémoriels se dessinent donc dans le corpus global : d'un côté, anciens combattants et souvenir et, de l'autre, mémoire et commémoration. Cette répartition vérifie les résultats obtenus lors de la description quantitative des cooccurrences de mots-clés et ceux produits par Calliope, en début de période (1984 et 2000). L'année 2010 était caractérisée par une substitution réciproque de souvenir à commémoration : mémoire apparaissait davantage avec souvenir, et anciens combattants avec commémoration. Nous tenterons de voir, au moment de l'analyse disjointe des sous-corpus annuels, si un tel changement transparaît dans les résultats.

Nous allons désormais décrire plus précisément chaque classe, en en mettant en valeur les présences et absences significatives. De cette façon, nous pourrons mieux cerner ce qui constitue la spécificité des classes, en montrant qu'une classe se détermine tout autant par la cooccurrence de ses termes que par leur distance aux termes des autres classes.
Précisons néanmoins que les différentes classes telles que présentées ci-dessous distinguent des univers discursifs différents mais qui peuvent apparaître au sein d'une même déclaration, d'un même texte. Si Alceste découpe un corpus en types de discours, ceux-ci s'entremêlent tout au long du corpus. En cela, la classification établie par Alceste ne saurait être confondue avec une typologie des documents traités.

3.1.1 Classe 1 : Les anciens combattants d'Afrique du Nord

Comme nous venons de le voir, la classe la plus homogène et la plus importante – en termes de couverture du corpus analysé –, rassemble les anciens combattants, avec un accent mis sur ceux d'Afrique du Nord. Leur vocabulaire est à la fois le plus et le plus important du corpus. Il désigne principalement, au sein des déclarations, les combattants et prisonniers qui ont participé, entre 1952 et 1962, à la guerre d'Algérie et aux combats du Maroc et de Tunisie Ces déclarations sont, pour la plupart, placées sous l'égide la Fédération nationale des anciens combattants en Algérie, Maroc et Tunisie (Fnaca[207]). Cette association se

[205] Le terme « hapax » désigne l'occurrence unique d'un terme à l'intérieur d'un corpus.
[206] Par richesse du vocabulaire, il faut entendre la part des mots pleins du corpus, en pourcentage. Le chiffre de 99,65% confirme l'adéquation de notre corpus aux hypothèses linguistiques postulées par Alceste, à savoir la prééminence accordée aux mots pleins dans une analyse textuelle.
[207] Pour en savoir plus sur les missions de la Fnaca, voir : http://www.fnaca.org/

distingue d'autres types d'associations d'anciens combattants (telles que l'UNC) en ce que son périmètre est précisément délimité à certains combats, ayant eu lieu sur une aire géographique[208] et à une période historique déterminées. Le caractère « récent » du conflit peut-il rendre compte de l'importance du nombre de déclarations, les combattants étant encore en vie ?
Le vocabulaire de cette classe fournit, en quelque sorte, une désignation des acteurs des conflits et une description du type d'associations auquel ils adhèrent : se voit en effet définit une communauté d'« anciens combattants » d'Afrique du Nord (« Algérie », « Tunisie », « Maroc »), « victimes », « prisonniers » de « guerre » ainsi que leurs « veuves », communauté qui se réunit en « associations » et « fédérations », lesquelles se déclinent au niveau local ou « départemental », en « sections », « amicales » et autres « comités ». Ces associations « républicaines » œuvrent en faveur de la « paix »

Parmi les énoncés spécifiques de cette classe, nous trouvons :

Unité textuelle n° 2534 Khi2 = 33 Individu n° 1309 *ID_1309 *date_1984 *dpt_LoireA *type_ann_2
(l) (association) (federation) (departementale) de la (loire) inferieure de (l) (association) (republicaine) des (anciens) combattants et (victimes) de (guerre) (comite) (departemental) de (l) (association) (republicaine) des (anciens) combattants et (victimes) de (guerre).
Unité textuelle n° 3155 Khi2 = 33 Individu n° 1715 *ID_1715 *date_1984 *dpt_Oise *type_ann_2
(ancien) (titre) (association) (amicale) des (anciens) combattants (prisonniers) de (guerre) et combattants (d) (algerie) (tunisie) (maroc) de plailly mortefontaine (association) des (anciens) combattants (prisonniers) de (guerre) et combattants (d) (algerie) (tunisie) (maroc/)
Unité textuelle n° 3950 Khi2 = 33 Individu n° 2124 *ID_2124 *date_2010 *dpt_Sarthe *type_ann_1
(acpg_catm) (anciens) combattants (prisonniers) de guerre_combattants (algerie_tunisie_maroc) (grouper) tous les (anciens) combattants (prisonniers) de (guerre), les combattants d' (algerie), (tunisie) et (maroc).

De même, pour les segments répétés les plus caractéristiques : « anciens combattants », « Maroc et Tunisie », « en Algérie », « amicale d'anciens combattants », « comité local », « mort pour la France », « prisonnier de guerre », « victime de guerre », etc.
De manière inverse, les termes qui cooccurrent le moins avec ceux de la classe sont « mémoire », « perpétuer », « souvenir », « mort », « défendre », « intérêt », « culture », etc.
De fait, le vocabulaire de cette classe fonctionne à la manière d'un titre[209] d'association, qui désignerait la qualité de ses d'adhérents (être d'anciens combattants) – fonctionnement assez typique des associations d'anciens combattants, qui nomment dès leur titre le statut d'anciens combattants.

3.1.2 Classe 2 : Maintenir le lien entre acteurs de conflits passés et perpétuer leur souvenir

Par opposition à la classe 1, mais en lien avec elle[210], le discours de la classe 2 correspond davantage à l'objet d'une association. Il ne s'agit plus ici de désigner les membres de l'association mais de spécifier ce qui les réunit. Son vocabulaire se caractérise par les termes de « défendre », dans l'« intérêt » « supérieur » du « pays », les « intérêts moraux, sociaux et matériels » des « adhérents » et de leurs « ayants-droits », de « maintenir » les « liens » de « camaraderie », d'« amitié » et de « solidarité », de « perpétuer » le « souvenir » des « morts » pour la « France » et la « Patrie », etc.
Notons que c'est au sein de cette classe qu'apparaît le terme « souvenir », l'un des mots-clés de la question mémorielle qui nous préoccupe. Cette présence exprime le besoin de

[208] Ainsi, la présence de noms de lieux et de pays est très importante au sein de cette classe (khi2 de 478) et celle d'adjectifs et d'adverbes (241).
[209] Ce qui explique la très faible présence de verbes au sein de cette classe (khi2 de – 111).
[210] Dans la mesure où la classe 2 constitue une subdivision d'une classe générique rassemblant les acteurs des conflits passés (classes 1 et 2).

maintenir un lien entre générations passées et actuelles – non plus le lien unissant la communauté vivante des anciens combattants mais celui qui prolonge aujourd'hui, parmi les ayants-droits notamment, la présence de ceux qui ne sont plus, de ceux qui sont morts pour la Patrie
De manière significative, puisqu'il s'agit de définir les missions des associations, la part de verbes, d'adjectifs et d'adverbes[211] est bien plus importante que dans la classe 1.

Voici quelques exemples d'énoncés typiques de la classe 2 :

Unité textuelle n° 1368 Khi2 = 58 Individu n° 707 *ID_707 *date_2010 *dpt_Gironde *type_ann_1
(union) (nationale) des (combattants) de soulac_grayan_vensac, (unc) de soulac_grayan_vensac (maintenir) (dans) l' (interet) (superieur) du (pays), les (liens) de (camaraderie), d' (amitie) et de (solidarite) (qui) (existent) (entre) (ceux) (qui) (ont) (participe) a la (defense) de la (patrie) (defendre) les (interets) (moraux), (sociaux) et (materiels) de (ses) (adherents) et (leurs) ayant (droits) (perpetuer) en (france) (metropolitaine), (dans) les (dom) tom,
Unité textuelle n° 1358 Khi2 = 51 Individu n° 704 *ID_704 *date_2010 *dpt_Gironde *type_ann_1
(union) (nationale) des (combattants) de cantenac_margaux_labarde, (unc) de cantenac_margaux_labarde, (maintenir) (dans) l' (interet) (superieur) du (pays) les (liens) de (camaraderie), d' (amitie) et de (solidarite) (qui) (existent) (entre) (ceux) (qui) (ont) (participe) a la (defense) de la (patrie); (defendre) les (interets) (moraux), (sociaux) et (materiels) de (ses) (adherents) et de (leurs) (ayants) (droit);
Unité textuelle n° 2090 Khi2 = 47 Individu n° 1031 *ID_1031 *date_2000 *dpt_IlleVilaine *type_ann_1
(union) (nationale) des (combattants) de guignen (maintenir) les (liens) de (camaraderie), d' (amitie) et de (solidarite) (entre) tous (ceux) (qui) (defendent) la (patrie); (defendre) les (interets) (moraux) et (sociaux) des (adherents) et de (leurs) (ayants) (droit), (perpetuer) le (souvenir) des (combattants) (morts) (pour) la (france), (developper) des (relations) avec des anciens (combattants) d' autres (nations);

Sont comparativement absents de cette classe les termes de « guerre », d'« association », d'« anciens », de « victimes », de « comités », d'« Algérie », de « Maroc » et de « Tunisie » qui appartiennent à la classe 1 et de « culture », d'« histoire » et de « patrimoine » qui relèvent de la classe 4.

Les deux discours de la première branche de l'arbre hiérarchique sont donc, bien que dissemblables dans leurs énoncés, fortement liés, à la manière dont un titre et un objet d'association le sont : certains individus (les anciens combattants de la classe 1) se réunissent en association, dans le but de maintenir les liens qui les unissent, de défendre leurs intérêts et de perpétuer le souvenir des morts pour la Patrie (classe 2).

3.1.3 Classe 3 : Rendre hommage aux victimes de discriminations

Cette classe se distingue des classes précédentes en ce qu'elle admet une communauté plus large que celle des anciens combattants et une mission qui ne consiste plus à en perpétuer le souvenir. Bien que se rattachant à la sphère de conflits passés, la classe 3 se situe pourtant du côté « culturel » de l'arbre de classification. La raison de ce positionnement s'explique par l'attitude adoptée vis-à-vis des conflits : il s'agit non plus d'une simple désignation des acteurs des conflits passés (classe 1), ni même du souhait d'en maintenir le souvenir vivant (classe 2), mais bien de rendre hommage aux victimes de conflits violents. C'est précisément cette dimension symbolique attachée aux acteurs et à leur environnement historique qui caractérise la classe 3. Une certaine forme de dramatisation et de mise en scène des conflits et de leurs protagonistes apparaît ici, qui faisait défaut aux classes précédentes.

[211] Les adjectifs et adverbes de cette classe ont un khi2 de 567, les verbes de 10. Ces particularités s'opposent de façon nette avec celles de la classe 1.

Il s'agit désormais d'« honorer », de rendre « hommage », via l'organisation de « cérémonies » « commémoratives » et l'érection de « monuments », aux victimes de conflits et de « discriminations » « idéologiquement » qualifiés (« nazisme », « fascisme », « barbarie », « génocide », « racisme », « antisémitisme »), pour mieux en souligner la dimension valeureuse (« résistance », « maquis ») et « sacrificielle ».

Unité textuelle n° 184 Khi2 = 72 Individu n° 90 *ID_90 *date_2010 *dpt_AlpesM *type_ann_2
(lutter) (contre) (toute) forme de (discrimination); realiser par l' union des (hommes) et des (femmes) de toutes (opinions) et (confessions), le (rapprochement) des (peuples) et l' (egalite) (parmi) les (etres) humaines, dans un (esprit) de (tolerance), de (fraternite) et de (respect) de la (laicite);
Unité textuelle n° 4414 Khi2 = 65 Individu n° 2383 *ID_2383 *date_2010 *dpt_ValOise *type_ann_1
en (erigeant) des (steles) ou (monuments) et notamment a l' occasion de (ceremonies) (commemoratives), (appel) du 18 (juin) sur la bbc a londres; (deces) du (general) de (gaulle) a colombey les deux (eglises) le 9 (novembre) 1970;
Unité textuelle n° 582 Khi2 = 66 Individu n° 292 *ID_292 *date_2000 *dpt_BouchesR *type_ann_1
defendre les victimes individuelles ou collectives du (racisme) et de l' (antisemitisme); (lutter) (contre) (toute) forme de (discrimination); realiser par l' union des (hommes) et des (femmes) de toutes (opinions) et (confessions) le (rapprochement) des (peuples) et l' (egalite) (parmi) les (etres) humains dans un (esprit) de (tolerance), de (fraternite) et de (respect) de la (laicite);

Le lien que cette classe établit au passé est un lien qui entend rassembler les individus au sein d'une mémoire collective qui dénonce les violences et les discriminations subies par les protagonistes. La commémoration (qui un l'un de nos mots-clés) offre alors le moyen de conserver et de diffuser la conscience, parfois nationale, d'un évènement de l'histoire collective, tout en servant d'exemple et de modèle.

S'adressant à une large audience – voire, de manière abstraite, à l'être humain, dans sa dimension universelle –, les événements dont il s'agit de se remémorer, au sein de la classe 3, ne sont pas géographiquement situés[212], contrairement à ceux de la classe 1, qui en rassemblait les acteurs. C'est bien davantage le contexte idéologique ayant généré des conflits (ou qui en génèrent encore) qu'il s'agit de rappeler que le lieu de tel événement particulier. A l'inverse, les dates[213] ont une présence significative et parsèment les énoncés associés à la classe 3 (nous en avons un exemple ci-dessous, avec le mois de « juillet »). L'inscription temporelle d'un événement, tout en lui procurant sa profondeur historique, en permet également le retour. Se dessine ainsi un calendrier des événements à célébrer, à commémorer, qui rythme l'histoire humaine (passée, présente et future) et vient rappeler, à chaque retour de la date anniversaire, le lien avec l'événement originel.

Concordancier de **juillet**dans le corpus

	appel du 18 juin, 14	**juillet**, 11 novembre, 5 decembre, ces prestations donneront leiur
les maquisards tombes face a l ennemi les 30 juin et 7	**juillet**	1944 entretenir le monument erige en juillet 1945 pour com
uin et 7 juillet 1944 entretenir le monument erige en	**juillet**	1945 pour commemorer leur souvenir comite du souvenir du m
stations patriotiques officielles, 8 mai, 18 juin, 14	**juillet**,	11 novembre et 5 decembre, et aux ceremonie de presentati
fetes nationales; vin d' honneur; don a la fete du 14	**juillet**	telethon.
ration du conseil municipal de valensole en date du 9	**juillet**	2004.
e territoire de la commune d' auxon, notamment les 14	**juillet**,	1er et 11 novembre, 8 mai, 18 juin,
ceremonies diverses lors-des journees souvenirs, 14	**juillet**,	11 novembre, etc, visites aux hopitaux des membres de l'
association des rescapes et victimes du 5	**juillet**	1962 a oran sauvagarde de la memoire du massacre perpetre
an sauvagarde de la memoire du massacre perpetre le 5	**juillet**	1962 a oran; organisation et manifestations sur sol nation

De manière significative, nous notons l'absence relative de la classe 3 de termes issus des classes 1 et 2 (combattant, Algérie, Tunisie, Maroc, solidarité, camaraderie, patrie, mort, guerre, etc.). La notion de combat apparaît sous la forme de nombreux segments répétés (combattre les idéologies, les falsifications, les discriminations, l'isolement, etc.), mais l'accent est mis ici sur la qualité militante des acteurs, leur lutte idéologique (via le verbe

[212] Le khi2 des noms de lieux et de pays est significativement de – 95, alors que, pour rappel, celui de la classe 1 était de 478.
[213] Les mois et jours ont un khi2 de 291 dans cette partie du corpus.

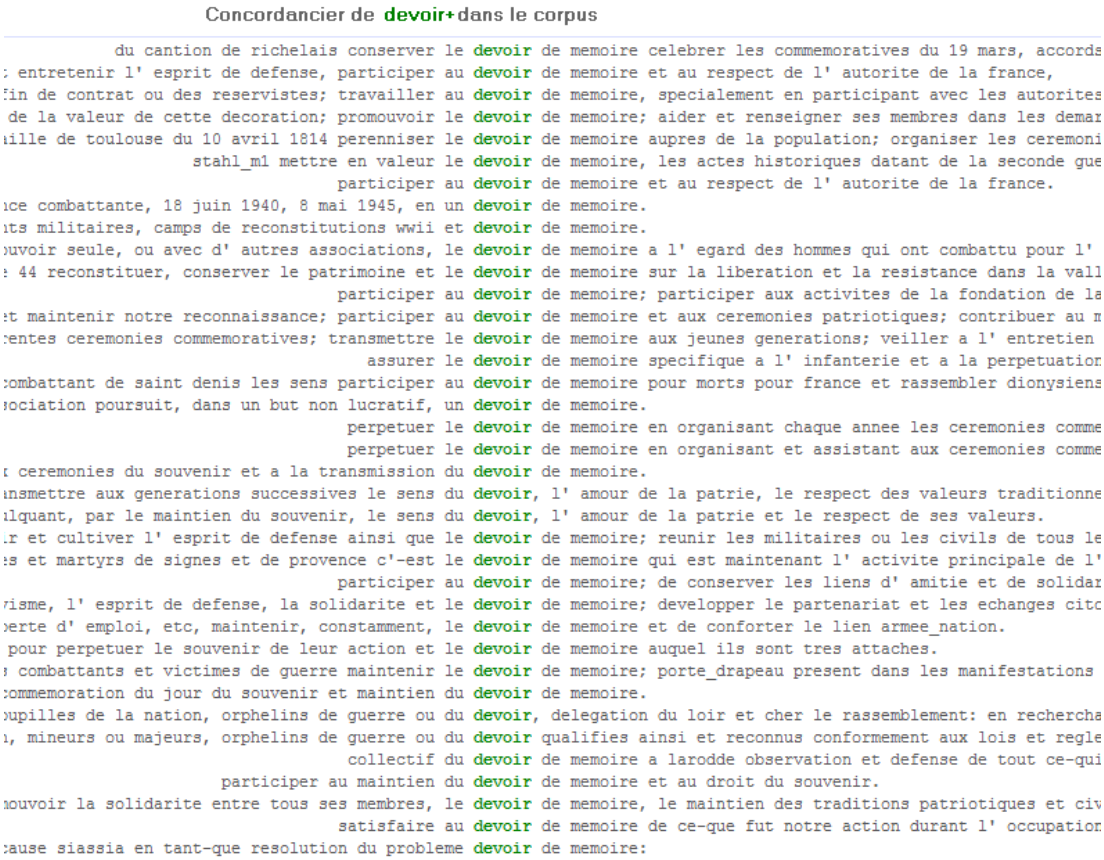
« combattre ») qu'il s'agit d'honorer, et non sur leur statut de « combattants », membres d'une armée régulière.

Par ailleurs, le terme « héros » n'est pas associé de façon significative à cette classe (ni à aucune autre classe du corpus, d'ailleurs) alors que l'on aurait pu l'attendre au sein d'une classe célébrant les individus ayant lutté contre les injustices, ou dans la classe 2, pour en perpétuer le souvenir. Le terme apparaît bien dans le corpus mais sa très faible fréquence, (11 occurrences) associée à son manque de spécificité (faible cooccurrence), expliquent qu'il ne soit pas jugé représentatif d'une classe. Il en est de même pour le verbe « célébrer[214] ».

Signalons une dernière thématique qui caractérise relativement plus la classe 3, mais de manière bien moins spécifique que celle de commémoration : il s'agit de l'orientation sociale de l'association, exprimée par les termes « réinsertion », « emploi », logement », « famille », « besoin ». Il s'agit, pour les membres de l'association, de « venir en aide » à leurs « compagnons blessés », de les soutenir dans une démarche de réinsertion sociale, de recherche d'emploi et de logement. Après un retour aux textes, il s'avère que ces termes proviennent tous de déclarations issues de la Fédération nationale des anciens des missions extérieures. Il faudrait pouvoir expliquer, dans une analyse future, pourquoi ces segments n'appartiennent pas à la classe 2 de l'arbre, qui traite de défense des droits et des intérêts des adhérents, de solidarité entre les membres de l'association et de leurs ayants-droits.

Par ailleurs, il est intéressant de noter que l'expression « devoir de mémoire » est considérée comme relevant plus spécifiquement de la classe 3, à la dimension symbolique et collective. Si l'expression « devoir de mémoire » apparaît également dans des énoncés relevant d'autres classes (15,63% de ses occurrences figurent dans la classe « mémoire », 14,84% dans la classe « anciens combattants » et 14% dans la classe « échanges »), elle caractérise principalement la classe « commémoration » (qui rassemble 28,13% de ses occurrences), en raison de sa cooccurrence plus importante avec les autre termes de cette classe. La classe « souvenir » constitue, quant à elle, le type de discours le moins en lien avec le « devoir de mémoire », ne regroupant que 1,56% de ses occurrences. Le devoir de mémoire implique un rapport au temps différent, qui se traduit par un public plus large que la perpétuation du souvenir des combattants morts, au sein des membres d'associations d'anciens combattants. En découlent aussi les moyens utilisés pour maintenir ce lien au passé : les cérémonies commémoratives. Le passé ne doit pas être oublié, ses exactions rappelées et ses victimes célébrées.

[214] Rappelons que le terme de « célébration » n'a pas été retenu comme mot-clé de filtrage des déclarations pertinentes. Si le verbe se retrouve parfois au sein de déclarations indexées par d'autres mots-clés, son absence de typicité, au regard de la thématique mémorielle, confirme l'analyse qui avait présidé à son éviction du corpus : l'acte de célébrer un événement ou une personne n'est que très faiblement associé aux domaines de la mémoire qui nous intéresse ici.



Mais l'expression « devoir de mémoire », bien que moins typique des autres classes, y apparaît néanmoins. Elle qualifie, dans les fragments de la classe 4, la dimension patrimoniale et historique propre à cette classe (« maintenir les valeurs et traditions marines, perpétuer le devoir de mémoire » ; « alliance internationale des villes pour le devoir de mémoire et le développement, favoriser les travaux de recherche et vulgarisation d'une histoire commune » ; « œuvrer pour préserver le devoir de mémoire et la transmission de notre histoire aux jeunes générations » ; « œuvrer pour la recherche et la défense de la vérité historique ainsi que pour le maintien du devoir de mémoire »).
Dans la classe 5, elle cooccurre avec des termes en liens aux liens et échanges, aux manifestations, fêtes et sports, spécifiques de cette classe (« regrouper les jeunes citoyens actifs dans le devoir de mémoire » ; « assumer, par le biais de manifestations festives et culturelles, le devoir de mémoire » ; « supporters du club de Montpellier Basket dans un but de convivialité et de devoir de mémoire » ; « développer toute action visant le devoir de mémoire des cyclistes martiniquais décédés »). A noter que cette expression figure au sein d'une déclaration d'association d'anciens combattants relevant de la classe 5, en lien donc avec l'aspect événementiel caractéristique de cette classe (« rassembler des anciens combattants, perpétuer le devoir de mémoire, organiser des manifestations cultuelles, culturelles et ludiques »).
Au sein de la classe 1, le « devoir de mémoire » concerne les associations d'« anciens combattants », souvent géographiquement situées : « anciens combattants d'Algérie, de Tunisie et Maroc TOE, section de Meilhan-sur-Garonne, la défense des droits et le devoir de mémoire » ; « associations d' anciens combattants, de résistants, de victimes de guerre, et garante du devoir de mémoire de la ville de Bron » ; « amicale des anciens combattants UNC et des amis du devoir de mémoire de Colleville-Montgomery » ; « regrouper tous les anciens combattants et victimes de guerre, entretenir le devoir de mémoire en organisant toutes les manifestations du souvenir » ; « regrouper tous les anciens combattants, leur apporter un soutien moral et matériel, ainsi qu' à leur veuve, maintenir le devoir de mémoire » ; « amicale des anciens combattants et sympathisants de Cornille, contribuer au devoir de mémoire, commémorer les armistices ».

3.1.4 Classe 4 : Conserver et valoriser le passé, la mémoire en héritage

La classe 4 du corpus fait état d'une dimension patrimoniale, caractéristique, du passé. Le passé devient ici un héritage à entretenir, qui met en lumière la richesse des productions et des activités humaines antérieures.
Il s'agit surtout de « connaître » et d'étudier un « patrimoine » (matériel ou immatériel), d'en transmettre la « mémoire » au moyen de « recherches » et d'« études » « scientifiques », d'en « préserver » et « valoriser » la dimension historique.
Quel que soit le domaine concerné (histoire d'un quartier, d'un village, d'une ville ; histoire culturelle ; histoire de techniques ; histoire sociale, etc.), des méthodes de conservation sont sollicitées (« archive ») visant à « sauvegarder », voire à « restaurer », un patrimoine constitué notamment de « documents oraux, écrits et photographiques », de « documents audiovisuels ». Ces traces matérielles du passé « témoignent » d'histoires humaines à transmettre aux générations futures : elles constituent ainsi un trait d'union entre passé et avenir, une mémoire à recueillir, conserver et léguer.

Les actions à entreprendre pour conserver ce patrimoine se traduisent par une présence plus significative des verbes[215] au sein des énoncés, relativement aux autres classes (comme valoriser, connaître, préserver, sauvegarder, rechercher, collecter, constituer, documenter, transmettre, découvrir, etc.), ainsi que celle des marqueurs d'intensité (plus, mieux, moins, etc.) qui qualifient l'ampleur de l'action ou du public concernés (« mieux [faire] connaître », « mieux comprendre », « porter à la connaissance du plus grand nombre », « public le plus large », « utilisation la plus complète », etc.) – comme s'il s'agissait de tenter de capturer et de transmettre le plus d'éléments possible sur un passé qui disparaît chaque jour un peu plus.

Unité texuelle n° 2435 Khi2 = 56 Individu n° 1236 *ID_1236 *date_2010 *dpt_LoireA *type_ann_1
reunir, (transmettre), (sauvegarder), (restaurer) (et) (mettre) (en) (valeur) (tous) les (elements) de la (memoire) des (transports) (en) (commun) de l' agglomeration nantaise: (photos), (films), (archives) administratives, (livres), (revues) (et) (journaux) traitant du (sujet), souvenirs (oraux) retranscrits, (petits) materiels, pinces, oblitereteurs, tickets, machines delivrant des tickets, etc, tenus, maquettes, materiels roulants;
Unité texuelle n° 568 Khi2 = 51 Individu n° 283 *ID_283 *date_2000 *dpt_BouchesR *type_ann_1
(collecter) (et) (diffuser) (tout) (ce-qui) (constitue) un (apport) (a) la (memoire), (a) l' (histoire) (et) au (patrimoine) (humains), (temoignages) (oraux), (ecrits), (images).
Unité texuelle n° 2542 Khi2 = 48 Individu n° 1315 *ID_1315 *date_2000 *dpt_Lot *type_ann_1
(institut) lotois, cgt, d' (histoire) sociale entretenir la (memoire) (collective); (laisser) un (temoignage) aux (generations) (futures); (mettre) (en) evidence dans l' (histoire) le (role) du (mouvement) (ouvrier); relier le (passe) au (present).

Le lien au passé se transforme, dans cette classe, en mémoire positive, à la fois individuelle et collective. Il ne s'agit plus de perpétuer le souvenir de personnes décédées aux combats ni de commémorer des événements tragiques mais, de manière plus large et plus affirmative, de collecter les traces, les empreintes, les témoignages d'activités humaines antérieures, afin d'en transmettre la valeur aux générations actuelles et futures. Le patrimoine culturel, social, historique d'une communauté devient un héritage vivant mais fragile, à entretenir.
Cet héritage est multiple, comme en témoigne la diversité des contextes dans lesquels apparaît le terme « mémoire » :

[215] Les verbes ont un khi2 de 55.

A noter également que c'est au sein de cette classe qu'est associée le plus spécifiquement l'expression « lieu de mémoire », ancrant, au moyen d'une inscription matérielle, géographiquement située, la dimension souvent évanescente, immatérielle de la mémoire. Mais cette expression apparaît également au sein de fragments de la classe 5 pour désigner des lieux de rencontre et d'échange autour d'une thématique (« par la création de lieux de mémoire : érection de monuments, apposition de plaques, ouverture d'un musée, d'une bibliothèque destinée au grand public », « création de jardins pédagogiques, botaniques, artistiques, de jardins, lieux de rencontre et lieux de mémoire ; organisation de visite; élaboration et diffusion d'informations sur la plante »). Le lieu de mémoire cooccurre, au sein de la classe 3, avec les notions de lutte contre les violence et de sacrifice (« lutter contre les idéologies d'inspirations fascistes, œuvrer à l'édification et à la préservation des lieux de mémoire » ; « défense des droits de l' homme et de la paix, elle œuvre à l'édification et à la prévention des lieux de mémoire » ; « création au 14 rue de paradis à Paris, d'un lieu de mémoire dédié au résistants juifs de la MOI en France sous l'occupation nazie en 1940 et 1945 »). Dans la classe 1, elle est associée aux anciens combattants (« ceux qui sont morts pour la France, combattants et victimes de guerre, faire respecter les lieux de mémoire »).

Les notions en lien aux conflits (anciens combattants, mort, guerre, victime, etc.) qui décrivaient les classes précédentes sont significativement moins associées à cette classe. L'aspect positif de la mémoire est ici affirmé, constitutif de la richesse du patrimoine humain.

3.1.5 Classe 5 : Favoriser les rencontres et les échanges

La dernière classe élaborée par Alceste regroupe des énoncés traitant des notions de « rencontres », d'« échanges », de « débats ». Cette classe se distingue par sa focalisation sur les liens sociaux, sur les liens interpersonnels, qui prime sur les domaines dans lesquels

ces liens sont exercés. En effet, si les secteurs d'activité concernés sont extrêmement diversifiés (pratiques artistiques, sportives, ludiques, activités pédagogiques, etc.) – à la différence des classes précédentes –, leur dénominateur commun est de « favoriser » les « échanges » et les « rencontres », par l'« organisation » de « manifestations », d'« animations », d'« ateliers », de « sorties », de « conférences » ou encore de « spectacles » et de « festival ».
Ce faisant, cette classe partage la dimension culturelle propre aux classes 3 (commémoration) et 4 (mémoire). Mais, à l'inverse de ces dernières, le lien qui unit les individus n'est plus un lien au passé mais un lien au présent, à soi, aux autres, qui se constitue dans l'exercice même des différentes pratiques, dans la mise en place d'événements, un lien qui n'existe que dans sa réalisation.
Outre la dimension culturelle propre à cette classe, des actions « pédagogiques » et sociales, en direction notamment des « jeunes » la spécifient. Il s'agit de « faciliter » l'« insertion » sociale, professionnelle de « personnes en difficulté », par la mise en place d'« ateliers ». Ici encore, l'importance du lien social transparaît. Ces éléments se différencient de la réinsertion qui apparaissait au sein de la classe « Commémoration », en ce qu'ils ne s'adressent pas aux mêmes personnes (il ne s'agit plus d'aider d'anciens combattants à se réinsérer dans la vie civile mais davantage de jeunes auxquels un « soutien scolaire » est nécessaire).

Unité textuelle n° 4002 Khi2 = 69 Individu n° 2156 *ID_2156 *date_2010 *dpt_SeineM *type_ann_1
(culture) co (organiser) et (proposer) des (manifestations) et (sorties) (culturelles) (principalement) au havre, dans sa (region) (mais) (aussi) en france et a l' (etranger); (accueillir), lors-de rencontres_debats, des (professionnels) de la (culture) et du tourisme; (promouvoir) la (formation) (universitaire) du master, memoire, mediation (culturelle) et tourisme (regional).
Unité textuelle n° 1564 Khi2 = 70 Individu n° 767 *ID_767 *date_2000 *dpt_Guadeloupe *type_ann_1
(organisation) de (sorties), (groupes) (musicaux) et de (danse), (theatre), moringue, (jeux) (divers), (voyages), (atelier) de travaux manuels, mise (en-place) de (soirees) a (themes), (expositions), (echanges) (culturels) entre (iles) et pays exterieurs;
Unité textuelle n° 3782 Khi2 = 66 Individu n° 2044 *ID_2044 *date_2010 *dpt_Réunion *type_ann_1
developper des (actions) d' (accompagnement) (scolaire) pour les (enfants) et (adolescents) des citees, des (ateliers) de communication lecture (ecriture) pour les (adultes), et illettrisme; (dynamiser) les quartiers (a-travers) des (ateliers) d' (activites) manuels, de (danse), d' (ecriture), (theatre), (chants), coutures, cuisines, musees, cinema, (randonnees), (excursions), pique_niques, (gymnastique), velo, (marche), (ou) bien des (sports) collectifs;

Etant donné la diversité des moyens d'expression employés dans les énoncés de cette classe ainsi que la multitude de leurs domaines d'application, une part significative est accordée aux noms[216], relativement aux autres classes.

Parce qu'ils sont éloignés de cette dimension sociale, qui vise un certain bien-être (collectif ou individuel) à l'occasion de moments partagés, de savoirs partagés, les lexiques des conflits passés apparaissent moins reliés au vocabulaire de la classe 5. Aussi les termes d'anciens combattants, de morts, de guerre, de souvenir, etc. sont-ils comparativement moins présents de cette classe.

Pour conclure, notons que cette classe est la plus éloignée de la thématique mémorielle au sens large, du lien au passé – d'ailleurs, aucun des mots-clés mémoriels n'y est significativement associé. Et lorsque ceux-ci apparaissent au sein des déclarations, les segments caractéristiques de cette classe n'expriment que le prisme culturel et événementiel du lien au passé, comme nous pouvons le constater dans les extraits ci-dessous :
Type de discours « souvenir » (vert) mêlé de termes de la classe 5 (« réunions », « sorties », en rose) :

[216] Les noms possèdent un khi2 de 43 dans cette catégorie de discours.

individu : 21 **** *ID_21 *date_2010 *dpt_Aisne *type_ann_1
association des anciens combattants et soldats nouvelle generation d' essigny le grand perpetuer le souvenir des combattants dont le sacrifice sur les divers champs de batailles, lors-des derniers conflits, a permis aux francais de vivre actuellement libres et egaux en droits, donner preuve de ce souvenir en participant avec ses membres aux diverses manifestations d' ordre national, lesquelles sont reconnues comme te, fete des deportes, 8 mai, appel du 18 juin, 14 juillet, 11 novembre, 5 decembre, ces prestations donneront leiur a defiler derriere les drapeaux des anciens combattants 1914_1918, 1939_ 1945, ffi, afn avec reunion au pied du monument aux morts de la commune, faire en sorte de creer un lien amical entre ses membres et de participer, s' il en est possible aux diverses reunion ou sorties organisees par l' association.

Type de discours « mémoire » (marron) intégrant une dimension festive ou sportive (en rose) :

individu : 225 **** *ID_225 *date_2010 *dpt_Aude *type_ann_1
lauragais au coeur organiser des actions a caractere culturel, festif ou sportif, visant a porter a la connaissance du grand public en-general, et des jeunes generations en-particulier, la croisade contre les albigeois, a commencer par les evenements de 1211, qui marquent un tournant dans l' histoire du lauragais; mobiliser le public de differentes communes autour de projets commemorant et valorisant les grandes dates historiques, les lieux de memoire et les faits culturels qui sont inscrits dans la memoire collective du pays lauragais.

Autre cas possible : l'emploi du terme « mémoire » au sens de « faculté cognitive » explique qu'il soit conservé comme typique de la classe 5 (bien-être, développement personnel, etc.) :

individu : 74 **** *ID_74 *date_2000 *dpt_Allier *type_ann_1
concordanse developpement et acces a tout un chacun de la pratique des activites gymniques et des techniques de detente et de relaxation; ses domaines d' application permettent, de-maniere generale, une meilleure connaissance de soi, ameliorent la concentration, la memoire, preparent aux examens ces differentes methodes utilisees s' ouvrent egalement a des publics cibles comme par-exemple aux sportifs, entrainement, concentration, gestion du stress.

Cette classe exprime donc la part « sociale » des déclarations qui peuvent, par ailleurs, entretenir un lien au passé (patrimonial, souvenir des anciens combattants) mais qui sera, quant à lui, formulé au moyen de termes relevant d'autres classes. Nous constatons donc bien ici le mélange de voix qui traversent le corpus, mis en avant par Alceste.

3.1.6 Analyse de la distribution des mots-clés

Afin de nuancer la présentation des classes que nous venons de faire, qui peut sembler figée, nous allons désormais étudier la distribution des mots-clés mémoriels, selon les différentes classes dans lesquelles ils apparaissent. En effet, l'élaboration des classes par Alceste se fonde sur l'analyse de fragments de texte et sur la cooccurrence des termes en leur sein. Il présente ensuite les termes qui sont les plus aptes à décrire une classe, en raison de leur cooccurrence plus importante avec les autres termes de cette classe qu'avec les termes d'autres classes. Il ne faut donc pas oublier qu'un terme peut apparaître au sein de différents types de discours, associés à différentes classes. Sa présence au sein de fragments de texte relevant de classes autres que celle de laquelle il est considéré comme le plus descriptif, présente donc un intérêt pour l'analyse : un terme présente autant de sens dans un fragment dans lequel il n'est pas significatif. Nous chercherons donc à voir comment les mots-clés mémoriels se déclinent dans les différents types de discours car, par l'analyse des divers contextes d'apparition, il est alors possible d'appréhender la polysémie des termes – ou comment un terme peut prendre une coloration différente en fonction du type de discours dans lequel il apparaît.

Distribution des mots-clés dans les classes[217]

	Classe1	Classe2	Classe3	Classe4	Classe5	**Nb total occurrences**
memoire+	69 5,45%	177 13,99%	187 14,78%	363 28,7%	152 12,02%	**1 265**
Commemorati+	36 16,07%	24 10,71%	64 28,57%	13 5,8%	36 16,07%	**224**

[217] Le tableau doit être lu comme suit : le premier chiffre (en noir) indique la fréquence du terme dans la classe, le second chiffre (en bleu) indique le rapport entre sa fréquence et le nombre total d'occurrences du terme au sein du corpus (en %). Le total de la fréquence dans les cinq classes ne correspond pas au nombre total d'occurrences des termes dans le corpus car certaines occurrences ne sont pas classées par Alceste.

Commemor+er	13 37,14%	1 2,86%	16 45,71%	1 2,86%	1 2,86%	**35**
Ancien	859 39,49%	180 8,28%	131 6,02%	48 2,21%	53 2,44%	**2 175**
Souvenir+	108 11,55%	494 52,83%	97 10,37%	68 7,27%	56 6%	**935**

3.1.6.1 Mémoire

Premier constat, le terme « mémoire », s'il caractérise davantage la classe 4 (« Patrimoine ») comme nous l'avons vu, apparaît néanmoins au sein d'autres classes et ce, de manière différenciée, en fonction du type de discours propre à chaque classe. Ainsi, lorsque les occurrences de mémoire (14%) apparaissent au sein de la classe 2 (« Souvenir »), elles désignent les combattants de manière générale (« servir la mémoire » des combattants morts pour la Patrie), tandis que lorsque mémoire figure dans la classe 1 (qui rassemble 5,45% de ses occurrences totales), c'est pour insister davantage sur le lieu des combats (« honorer la mémoire » des morts en Afrique du Nord) et sur le lien direct avec les anciens combattants (« entretenir la mémoire », de « défendre l'honneur et la mémoire », d'« entretenir le devoir de mémoire » des camarades morts en service, etc.).
Par ailleurs, près de 15% des occurrences de mémoire figurent dans la classe 3 (« Commémoration »), en lien notamment au « devoir de mémoire ». L'accent est mis ici sur la mémoire des victimes de discriminations et de ceux qui ont lutté (« honorer la mémoire » des victimes de persécution, des maquisards, des Français assassinés à Mauthausen, de « participer au rayonnement de la mémoire » des forces françaises libres).
Comme énoncé plus haut, les occurrences de mémoire (12%) dans la classe 5 (« Echanges ») prennent des sens différents : soit celui de faculté cognitive (« entretenir, stimuler la mémoire », « atelier de mémoire », « travail sur la mémoire », etc.), soit celui de publication (« élaboration de rapports de stage, de mémoires »), ou encore de « mémoire vive », au sens informatique. Lorsqu'il apparaît avec le sens de « mémoire collective », il peut désigner une manière d'entretenir le lien intergénérationnel (« insuffler une dynamique à la mémoire collective, pallier les isolements et la disparition des riches savoir-faire », « favoriser les échanges entre générations et la mémoire », « actions de culture de la mémoire envers les scolaires », « visites guidées proposées en particulier aux enfants à l'espace mémoire ». Il peut enfin désigner la finalité d'enregistrements audiovisuels (« constituer une mémoire vidéo comme support de réflexion »).

3.1.6.2 Commémoration

Si le terme commémoration, et le terme apparenté (commémoratives), sont principalement rattachés à la classe 3 par Alceste (en raison du nombre d'unités de la classe (28,57%) contenant ces deux termes, en lien à d'autres termes des fragments), il faut souligner qu'ils apparaissent fréquemment au sein des classes 5 et 1 (16% de leurs occurrences totales dans chaque classe). S'ils qualifient les cérémonies d'hommage aux victimes de discriminations, dans la classe 3, que représentent ces termes dans les autres classes ?
De manière typique de la classe 5, ils désignent des cérémonies favorisant les échanges interculturels et s'inscrivent dans un ensemble de manifestations événementielles (« enseignement des traditions, de la langue et la commémoration des fêtes », « favoriser l'échange interculturel, actions ponctuelles, commémoratives, culturelles » ; « manifestations musicales populaires, en prêtant son concours aux commémorations et aux fêtes organisées par la ville » ; « organiser des activités culturelles, rencontres, colloques, commémorations, visite des lieux » ; « animer des activités culturelles, sportives ou commémoratives »).

Au sein de la classe 1 (« Anciens combattants »), ils désignent principalement les commémorations du cessez-le-feu en Algérie du 19 mars 1962 ou les anciens combattants d'Afrique du Nord (« commémoration annuelle de l'anniversaire du cessez-le-feu en Algérie le 19 mars 1962 » ; « œuvrer en faveur de la paix, notamment par la commémoration du 19 mars 1962 » ; « organiser en commun les cérémonies commémoratives d'anciens combattants prisonniers de guerre et combattants en Algérie, Tunisie, Maroc ») ou pour réunir d'anciens combattants (« retrouvailles d'anciens camarades parachutistes militaires et commémoration de notre fête la Saint-Michel »).
Dans la classe 2 (qui rassemble près de 12% des occurrences), la commémoration participe de la célébration d'anciens combattants (« développer des relations fraternelles entre les anciens combattants des nations amies ou alliées, en organisant des cérémonies commémoratives, patriotiques ou religieuses » ; « perpétuer le souvenir des combattants morts pour la France, servir leur mémoire en organisant des cérémonies commémoratives »).
Dans la classe 4 (classe la moins concernée par les commémorations, avec moins de 6% d'occurrences), enfin, les cérémonies commémoratives sont associées à une dimension historique et patrimoniale, que ce soit pour célébrer divers événements ou personnages historiques (« situer cet héritage historique dans le cadre de la commémoration du millénaire » ; « révolution française de 1789 auprès du public et contribuer à la commémoration de son bicentenaire » ; « cette commémoration rappelle l'histoire de Siméon le Stylite » ; « soutenir le chef de la maison de Bourbon dans ses entreprises sociales culturelles, caritatives, patrimoniales, commémoratives et historiques ») ou pour traiter de l'histoire propre aux anciens combattants et du patrimoine militaire (« anciens combattants d'Autheuil-Maletable, commémorations et manifestations, publications liées a l'histoire des anciens combattants » ; « conservation de véhicules militaires et d'équipements de type abandonnés et organiser des manifestations commémoratives historiques » ; « commémorations été 44, préservation de tous types de véhicules militaires »).

A noter également que le verbe « commémorer » a été associé à la classe 3 par Alceste (tout comme les termes commémorations et commémoratives l'étaient) mais que sa fréquence d'apparition très élevée dans la classe 1 (37,14% de ses occurrences). S'il s'agit dans la classe 3 de commémorer le souvenir d'actions héroïques, de sacrifices ou de victimes de discriminations (« commémorer le souvenir des troupes françaises libres de la Libération » ; « commémorer leur souvenir, comité du souvenir du maquis de la Coupille » ; « commémorer le souvenir de la déportation des homosexuels mis à mort par les nazis en aout 1944 » ; « commémorer les faits d'armes de nos armées, honorer nos morts et perpétuer le souvenir des sacrifices » ; « commémorer l'armistice des guerres gestes respectueux aux souvenirs des disparus »), dans la classe 1, l'accent est mis sur les anciens combattants, avec souvent, la mention d'un lieu géographique (« commémorer les armistices des guerres et AFN, réunir les anciens combattants de Boisredon et Soubran » ; « perpétuer et commémorer le souvenir des anciens combattants » ; « œuvrer en faveur de la paix en commémorant la date du 19 mars 1962 » ; « associations d'anciens combattants du canton de Laignes et les aider dans leurs tâches visant à commémorer le souvenir de leurs camarades morts pour la France » ; « amicale des anciens combattants et sympathisants de Cornille, contribuer au devoir de mémoire, commémorer les armistices, resserrer les liens d'amitié entre anciens combattants » ; « association des anciens combattants de Chatain, section 19, commémorer et honorer les dates anniversaire des derniers conflits »).
Les autres classes sont très peu concernées par ce verbe, dans la mesure où une seule occurrence y figure à chaque fois. Dans la classe 2, il rappelle le souvenir de ceux qui ont « donné leur vie pour la Patrie » et, au sein de la classe 5, il est rattaché à des activités pédagogiques de transmission du passé (« éducation et tout autre type d'événement visant à commémorer et promouvoir les expériences passées »). A noter que la classe 4 est beaucoup moins associée au verbe commémorer qu'elle ne l'était avec les termes commémoration et commémoratives. Ce qui caractérise ce type de discours est donc davantage le fait d'organiser des cérémonies, notamment commémoratives, que l'action elle-

même de commémorer un événement. Il s'agit ici de « mobiliser le public de différentes communes autour de projets commémorant et valorisant les grandes dates historiques »

3.1.6.3 Anciens combattants

Si près de 40% des occurrences d'« Anciens combattants » relèvent de la classe 1, il faut souligner qu'une proportion plus importante (41,56% des occurrences) n'est pas classée par Alceste. Il faut préciser ici que ce n'est pas l'expression « anciens combattants » qui est analysée mais seulement le terme « ancien ». Il aurait fallu (et c'est précisément ce que nous ferons dans nos prochaines analyses) associer « anciens » et « combattants » au moyen d'un tiret bas pour que ces deux termes soient traités ensemble, sans distinction. C'est pourquoi, au sein des autres classes, moins marquées par la présence conjointe d'anciens combattants, le terme « ancien peut apparaître seul.
Dans la classe 1, il s'agit essentiellement de préciser l'ancrage local de l'association d'anciens combattants concernée (« association stéphanoise d'anciens combattants et victimes des guerres » ; « association de la fédération départementale de la Loire inférieure de l'association républicaine des anciens combattants » ; « amicale des anciens combattants prisonniers de guerre et combattants d'Algérie, Tunisie, Maroc de Plailly-Mortefontaine » ; « association républicaine des anciens combattants et victimes de guerre, ARAC section locale » ; « section d'Entraigues-sur-la-Sorgue de l'association républicaine des anciens combattants »).
Les autres apparitions du terme « ancien » se répartissent de la manière suivante : dans la classe 2, il s'agit de « développer des relations fraternelles entre les anciens combattants des nations unies ou alliées », dans la classe 3, de « faciliter l'obtention d'une carte et d'un statut d'ancien combattant, et de promouvoir l'esprit civique et la valeur morale ».
Dans les deux dernières classes, très peu de liens avec les associations d'anciens combattants sont entretenus, ce qui explique que le terme ancien soit employé dans des contextes dans lesquels combattants n'apparaît pas : ainsi, dans la classe 4, le terme ancien concerne la conservation d'un patrimoine (« favoriser la conservation des variétés végétales anciennes » ; « association de préservation du patrimoine local transformation de l'ancien presbytère de Remoray ») ; et enfin, dans la classe 5, le développement de relations au sein d'une communauté, d'anciens élèves notamment (« réseau relationnel, professionnels bénévoles, anciens élèves » ; « rencontres entre étudiants anciens et actuels »).

3.1.6.4 Souvenir

Nous avons vu que le mot-clé mémoriel « souvenir » était considéré par Alceste comme le plus descriptif de la classe 2 – ce qui correspond ici à sa fréquence d'apparition (cette classe rassemble en effet près de 53 % de ses occurrences totales). Il apparaît néanmoins dans la classe 1, pour 11,55% de ses occurrences. Dans ce contexte, il traite d'anciens combattants (« maintenir vivace le souvenir des anciens combattants » ; « participer aux cérémonies civils et militaires du souvenir avec les anciens combattants »). Dans la classe 3, il s'agit de « commémorer le « souvenir des troupes françaises libres de la Libération » et dans la classe 4, de la dimension patrimoniale et historique (« souvenirs d'Arveyres, promouvoir l'histoire de notre commune », « protection des monuments, sites et souvenirs »).

3.1.7 Spécificité des classes et variables illustratives : Répartition temporelle, géographique et par type d'annonce

Après avoir montré comment chaque classe se définissait en se distinguant des autres par le biais de la typicité de leurs énoncés, nous avons voulu savoir si ces classes étaient caractérisées par d'autres éléments que les éléments lexicaux, et notamment si certaines variables (temps, espace, type d'annonces) se retrouvaient plus spécifiquement associées à certaines classes. Autrement dit, nous avons tenté de voir comment se déclinait la question mémorielle, non plus en interne, au sein des univers de discours, mais à l'échelle du temps,

de l'espace ou en fonction du type d'annonce : une classe de discours est-elle surreprésentée (ou sous-représentée) à une date, dans un département, par un type d'annonce ?

Nous avons dressé un tableau résumant les relations entre variables (date, département, type d'annonce) et classes[218] :

	Date	**Département**	**Type d'annonce**
Classe 1 **Anciens combattants**	1984 (699) 2010 (- 207 2000 (- 44)	Charente (42) Var (22) Paris (- 63) Nord (- 56)	Modification (149) Création (- 143)
Classe 2 **Souvenir**	2000 (770) 2010 (- 269) 1984 (- 211)	Nord (274) Ille-et-Vilaine (253) Ardennes (236) Paris (- 83) Bouches-du-Rhône (- 27)	Création (101) Modification (- 105)
Classe 3 **Commémoration**	2010 (100) 2000 (- 58)	Alpes maritimes (45) Corrèze (14) Rhône (13) Nord (- 25)	Création (6)
Classe 4 **Mémoire**	2010 (162) 2000 (- 53) 1984 (- 45)	Paris (149) Aveyron (21) Pyrénées orientales (19) Ardennes (- 18) Gironde (- 15)	Ni création (0) ni modification (0)
Classe 5 **Echanges**	2010 (217) 2000 (- 90) 1984 (- 42)	Paris (62) Guadeloupe (48) Réunion (42) Ille-et-Vilaine (- 19)	Création (4)

Ce tableau nous fournit déjà une première indication quant à la typicité (ou la moindre typicité) d'une classe relativement à une variable. Une classe peut ainsi avoir une très forte probabilité d'être associée à la modalité d'une variable (date, département ou type d'annonce) et une probabilité négative plus ou moins forte d'être reliée à une autre modalité de la variable. Des probabilités très faibles voire nulles impliquent souvent l'absence de la probabilité inverse. Divers scénarios de corrélation se dessinent donc.

3.1.7.1 Répartitions temporelles

Une certaine ventilation temporelle des classes semble se profiler, avec un pôle plus ancien pour la première branche de la classification, qui se traduit par une prééminence du lien

[218] Les éléments de ce tableau, fournis par Alceste, sont obtenus par une mesure du khi2, qui calcule la probabilité d'appartenance d'une variable à une classe. Nous indiquons dans le tableau, en bleu et entre parenthèses, les valeurs des khi2 pour chaque modalité de la variable, et en vert les khi2 négatifs, c'est-à-dire les probabilités négatives pour une variable d'être associée à une classe.
Par manque de temps, n'entrent pas dans ce tableau d'analyses croisées entre plusieurs variables (date et département, par exemple). Les variables sont donc analysées ici indépendamment les unes des autres.

entre le discours des anciens combattants et l'année 1984 et entre celui du souvenir et l'année 2000. La seconde branche de la classification, qui englobe les déclarations dont les énoncés traitent de la mémoire, de la commémoration et des échanges, est davantage associée à l'année 2010. Cela signifie que ces années-là sont, à l'échelle du corpus, comparativement surreprésentées au sein de ces classes, au détriment des autres années. Notons que la corrélation entre classe et date est beaucoup plus élevée pour la classe des anciens combattants et celle du souvenir que les classes à coloration culturelle.

Nous retrouvons ces informations sous une autre forme, dans les graphiques ci-dessous, qui montrent les termes les plus représentatifs[219] par année :

Vocabulaire représentatif de l'année 1984

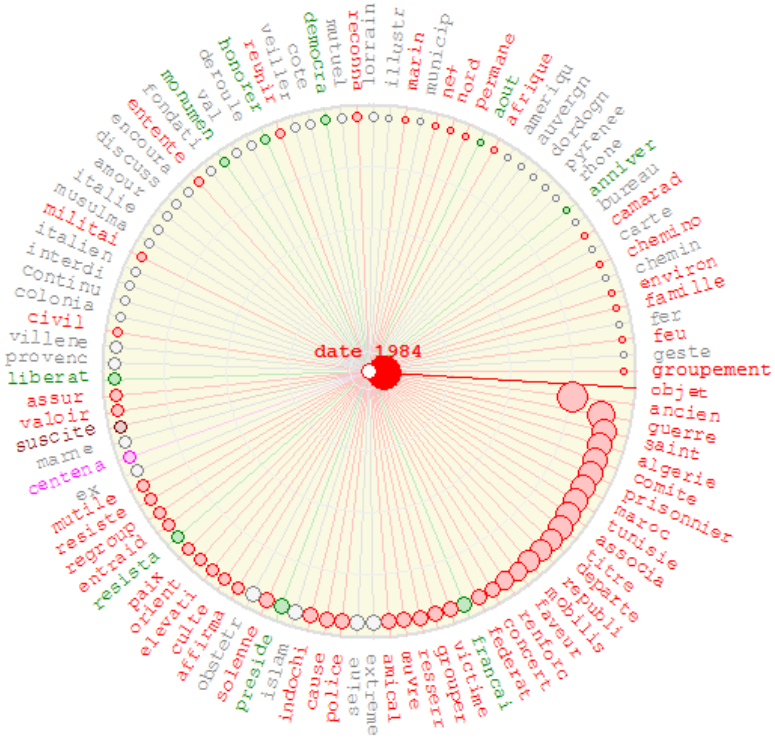


Figure 45 – Disque Année 1984 / Formes réduites

[219] La taille des pastilles est proportionnelle au poids des termes au sein de leur classe et leur distance au centre du disque indique le degré d'association du terme à la classe (plus la pastille est proche du centre, plus le terme est associé à la classe). La coloration des termes et des pastilles correspond à la couleur de leur classe d'appartenance : rouge = anciens combattants ; bleu = souvenir ; rose, marron et vert = commémoration, patrimoine et échanges.

Vocabulaire représentatif de l'année 2000

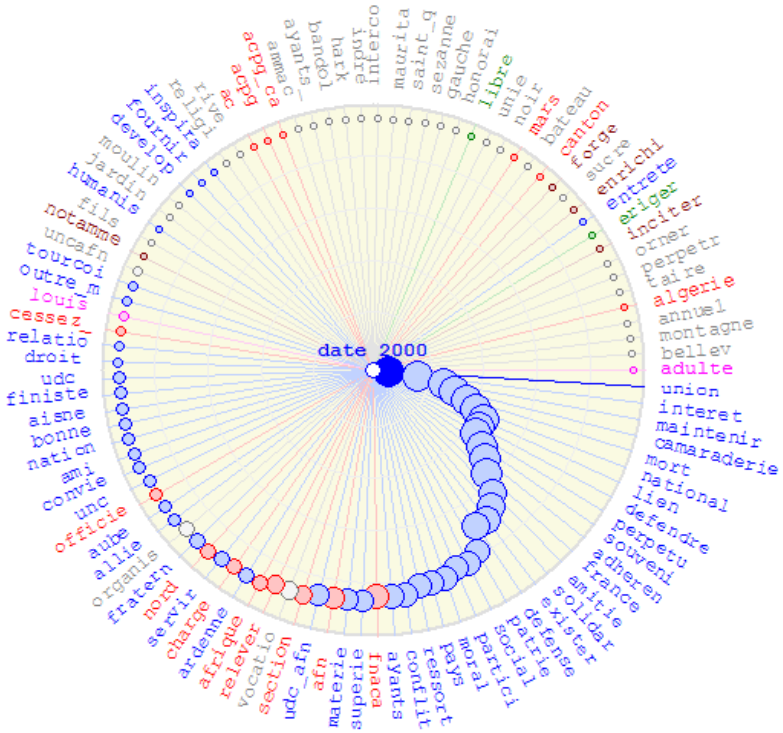


Figure 46 – Disque Année 2000 / Formes réduites

Vocabulaire représentatif de l'année 2010

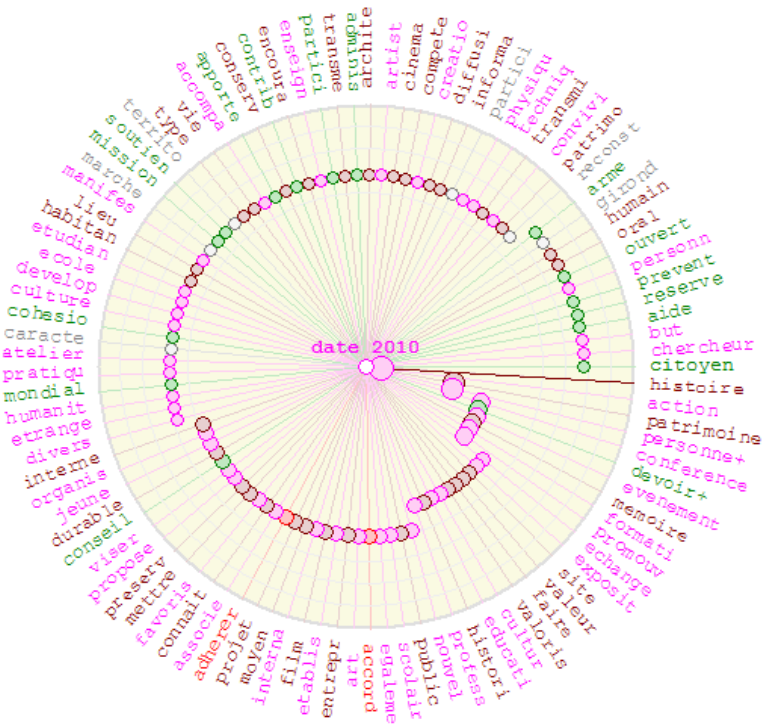


Figure 47 – Disque Année 2010 / Formes réduites

Chaque année du corpus est bien caractérisée par une classe : l'année 1984 par la classe des « Anciens combattants », 2000 par celle du « Souvenir » et 2010 par les trois classes de la branche « culture » (« Patrimoine », « Commémoration » et « Echanges »). Le type de structure de ces graphiques nous apprend en outre que les discours sont plus ou moins associés aux années. Ainsi, l'année 2000 est plus fortement spécifiée par les termes de la classe « Souvenir » que l'année 2010 ne l'est par les termes de la branche « culture ». La structure du disque de l'année 2000 est en effet plus concentrée et inégale (pastilles bleues proches du centre et de taille importante) que celle de l'année 2010, qui montre une forme plus diversifiée, plus également répartie (pastilles de différentes couleurs, ayant une taille similaire et distribuées de manière plus équilibrée autour de centre). L'année 1984, quant à elle, est caractérisée par la classe des anciens combattants, avec des termes au poids important mais qui lui sont moins fortement associés que ceux de la classe « souvenir » à l'année 2000. Nous en déduisons que les termes associés à 1984 doivent également l'être à d'autres années.
Ce que nous confirment les graphiques ci-dessous qui montrent la répartition des unités textuelles, selon les années, contenant les termes « prisonnier » (classe rouge « Anciens combattants ») et celles contenant le terme « mort » (classe bleu « Souvenir ») : les énoncés qui contiennent le terme « prisonnier » sont effectivement davantage répartis sur les trois années du corpus (mais ils caractérisent fortement la classe « Anciens combattants, car ils apparaissent quasi-exclusivement en rouge) que les énoncés contenant le terme « mort », qui spécifient beaucoup plus l'année 2000.

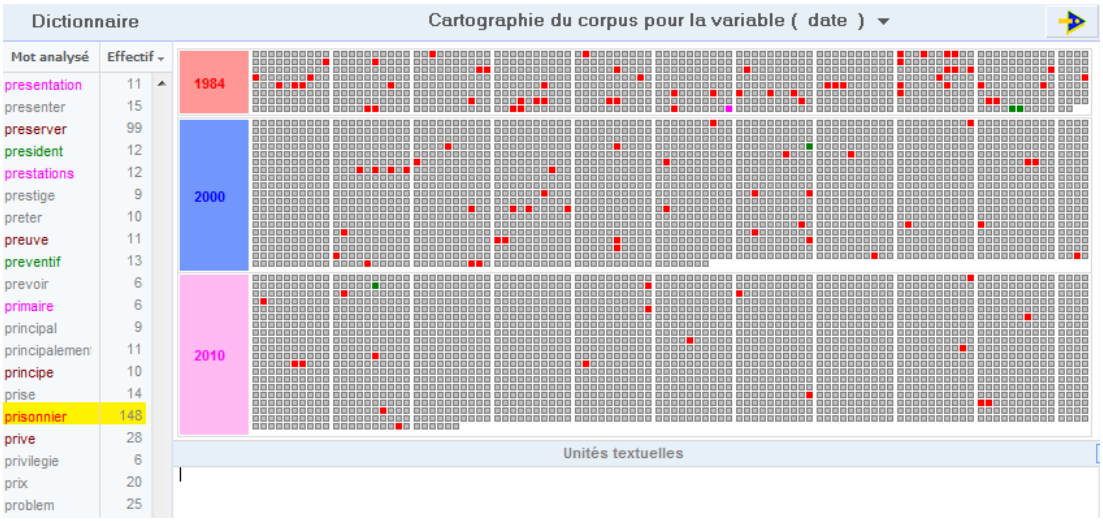


Figure 48 – Distribution du terme « prisonnier » par année

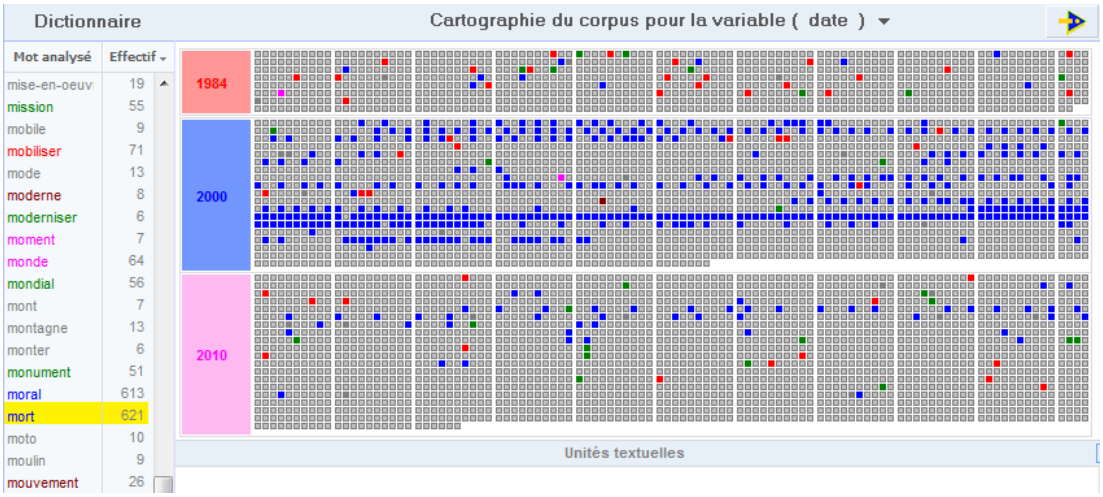


Figure 49 – Distribution du terme « mort » par année

Nous retrouvons également, dans le graphique ci-dessous, l'éparpillement relatif (ou la moindre spécificité) des termes qui caractérise l'année 2010 – par exemple, ici, le terme « devoir+ » (qui désigne le « devoir de mémoire »), s'il apparaît davantage au sein des énoncés de l'année 2010, apparaît également les autres années mais surtout, il apparaît au sein d'énoncés relevant de plusieurs classes (cases de différentes couleurs). Ce qui confirme l'information donnée par la taille et la distances des pastilles sur le graphique précédent. En cela, l'association des termes qui caractérisent l'année 2010 à des types de discours est moins forte relativement que celle des autres classes.

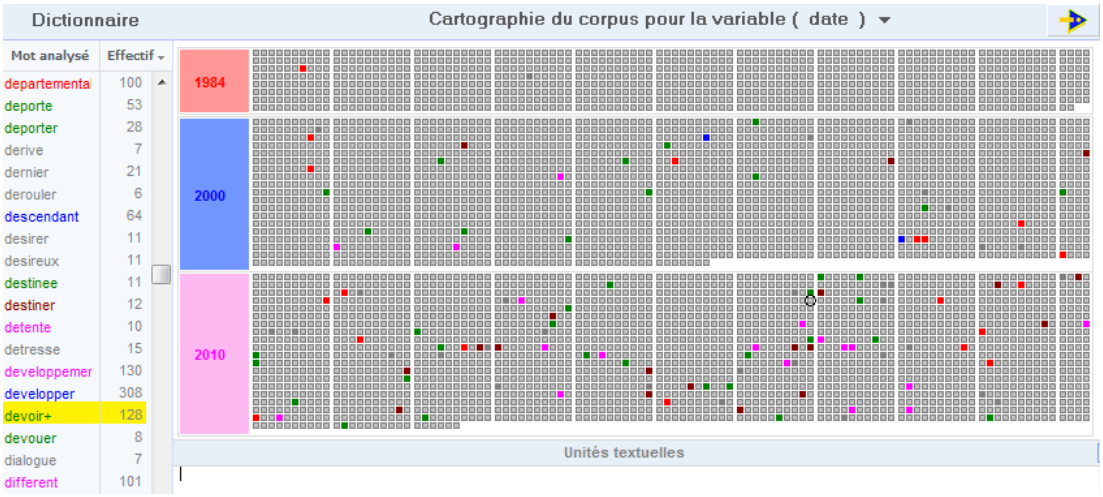


Figure 50 – Distribution du terme « devoir+ » par année

3.1.7.2 Répartitions administratives (type d'annonce)

La distribution temporelle des classes se voit, en quelque sorte, précisée par celle du type d'annonce. Ainsi, l'ancienneté des associations des anciens combattants, relativement aux classes de la branche « culture », est réaffirmée, voire accentuée, par le fait que leurs déclarations sont caractérisées par davantage de modifications que de créations. L'ancienneté de cette classe serait ainsi plus importante encore – ou plutôt le degré d'activité « administrative »[220] de ces associations est à la fois plus grand en 1984 et concerne surtout des modifications de déclarations ayant été créées antérieurement. La classe « souvenir » est, quant à elle, spécifiée par la catégorie « création », ce qui souligne une relative jeunesse de ce type de discours. Par contre, le lien entre discours et type d'annonces est, pour les classes à dominante culturelle, beaucoup plus faible, voire inexistant. Cette information ne constitue donc pas pour ce type de discours une variable discriminante.

3.1.7.3 Répartitions géographiques

Du point de vue de la distribution géographique des classes, nous constatons, là encore, que des relations spécifiques se dégager entre département et classe, qui double la structuration en deux grands types de domaines. Les discours sur le souvenir (classe 2) sont ainsi très fortement associés aux départements du Nord, de l'Ille-et-Vilaine et des Ardennes, à tel point que les quatre autres classes (anciens combattants, mémoire, commémoration et échanges) sont comparativement beaucoup associées à ces départements. La corrélation entre département et classe est bien moindre pour les anciens combattants, ce qui indique un éparpillement géographique.

[220] Par degré d'activité administrative, nous entendons l'activité des associations reflétée par les publications du JO.

Quelques termes se dégagent particulièrement sur le graphique ci-dessous (défendre, mort, perpétuer, union, France, souvenir, etc.) : ce sont les cooccurrences de ces termes au sein des énoncés relevant de la classe 2 qui spécifient le département du Nord.

Vocabulaire représentatif du département Nord[221]

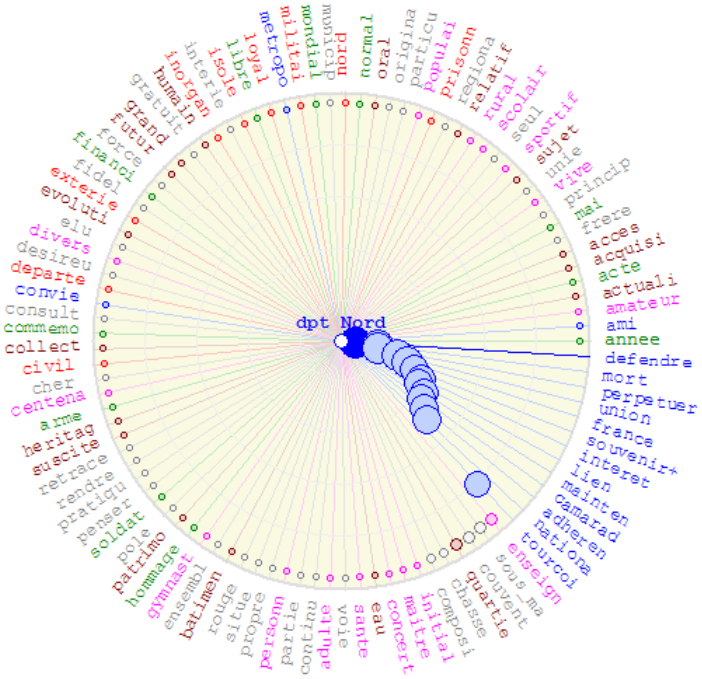


Figure 51 – Disque Département Nord / Formes réduites

De son côté, la branche « culture » (essentiellement, les classes « Patrimoine »et « Echanges ») est fortement liée à Paris. Mais la distribution des termes est moins concentrée que celle qui caractérisait le département du Nord. De même que pour l'année 2010, les énoncés de la branche « culture » associés à Paris relèvent de plusieurs classes (et non d'une seule, comme pour le département du Nord).
L'association qui se dessine entre Paris et la culture permet de déduire une certaine inégalité géographique en termes de biens culturels entre Paris et le reste de la France. Cette relation particulière entre les classes de la branche culturelle et Paris explique, par ailleurs, le lien relativement moins fort à ce lieu, dans un mouvement réciproque à celui mentionné plus haut, des classes liées aux anciens combattants et au souvenir.

[221] Le vocabulaire du « souvenir » lié au département du Nord est fortement concentré sur quelques termes, de poids très important. Celui lié à Paris, dans le second disque, est réparti de manière plus équilibrée et plus diversifiée.

Vocabulaire représentatif de Paris

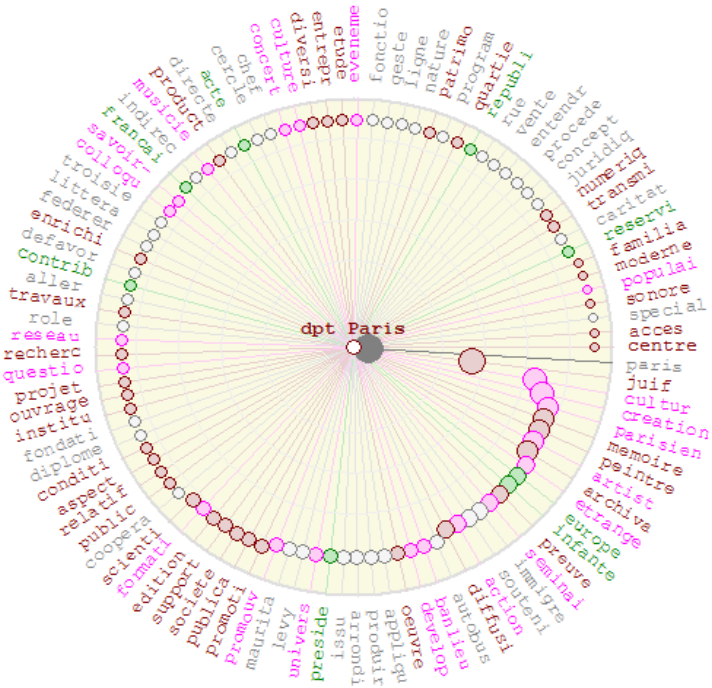


Figure 52 – Disque Département Paris / Formes réduites

Nous voyons donc des pôles émerger dans la distribution temporelle et spatiale notamment, de la question mémorielle. Ainsi, les déclarations relevant du domaine des anciens combattants semblent caractérisées par une certaine ancienneté (forte corrélation à l'année 1984 et au type « modification de l'association »), alors que celles du souvenir sont plus récentes (fort lien à l'année 2000 et au type « création d'association »). Les discours commémoratifs, mémoriels (au sens strict) ou traitant d'échanges sont encore plus jeunes (année 2010) mais sans que le type d'annonce soit caractéristique.
Enfin, une opposition très nette se dessine entre les discours sur le souvenir et la mémoire quant à leur ancrage territorial. A noter que la classe des anciens combattants est davantage définie par une corrélation négative quant à son appartenance géographique. En effet, si aucun département ne s'impose vraiment pour ce type de discours, le lien à Paris est comparativement plus faible. Les déclarations d'associations d'anciens combattants sont donc, sur la période étudiée, plutôt d'origine provinciale – ce qui est confirmé par la mention, dans le titre, du nom de nombreuses sections locales de fédérations nationales.
Ces informations présentent l'avantage, outre celui de pouvoir caractériser une classe au moyen de diverses variables, d'indiquer des points saillants d'une classe, qu'il s'agira, pour mieux en saisir la nature, d'étudier précisément par un retour averti aux textes et par une analyse de leur contexte d'apparition. Pour qui souhaite analyser finement le vocabulaire mémoriel au sens strict, ie le lexique employant le terme « mémoire », les déclarations enregistrées à la préfecture de police de Paris présentent un intérêt indéniable ; de même pour le lexique du souvenir et les départements du Nord, de l'Ille-et-Vilaine et des Ardennes.
Ces indications appellent par ailleurs à être affinées, par un élargissement par exemple de la taille de l'échantillon sur l'axe temporel. Comment évolue l'ensemble des discours avant 1984 ? Après 2010 ? Les dates non retenues dans cet échantillon confirment-elles ou infirment-elles les tendances ébauchées ici ? Les résultats exposés ici ne font que suggérer certaines pistes qu'il faudra poursuivre afin de pouvoir mieux appréhender le phénomène mémoriel.

3.2 Analyse chronologique des trois sous-corpus annuels

Après avoir traité en un corpus unique les fichiers des 3 années de déclarations au JO, nous allons désormais présenter les résultats issus de traitements individualisés. Plus précisément, nous nous concentrerons dans cette partie sur une comparaison entre les résultats issus des deux processus d'analyse, pour montrer ce qui est commun ou qui diffère. Nous ne décrirons donc pas finement chaque classe, année par année, mais seulement les similarités ou les divergences qui peuvent apparaître par rapport à l'analyse globale du corpus que nous avons effectuée précédemment. Nous chercherons, dans cette perspective chronologique, essentiellement à voir comment les types de discours évoluent d'une année à l'autre, ainsi que leur répartition spatiale et le type d'annonce qui les qualifie

	1984	**2000**	**2010**
Pourcentage du corpus analysé	82%	77%	87%
Nombre total d'occurrences	21 023	56 259	63 413
Nombre total de formes distinctes	2 776	4 802	6 839
Nombre de hapax	1 558	2 666	3 470
Richesse du vocabulaire	99%	99,52%	99,20%

D'une manière générale, et conformément aux résultats de l'analyse globale, les 3 corpus traités individuellement présentent un taux de pertinence du traitement élevé (> à 70%) et font état d'une grande richesse du vocabulaire – à entendre comme une forte proportion de mots pleins. Ceci n'est pas étonnant dans la mesure où il s'agit du même corpus, découpé en sous-parties.

Classification de l'année 1984

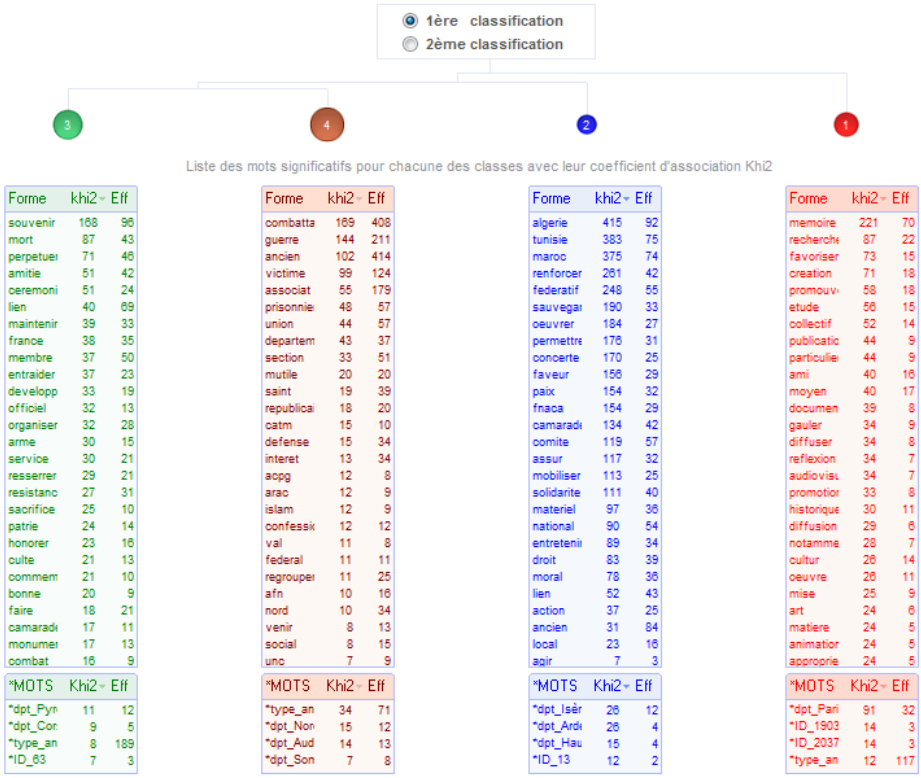


Figure 53 – Classification descendante hiérarchique de l'année 1984

Classification de l'année 2000

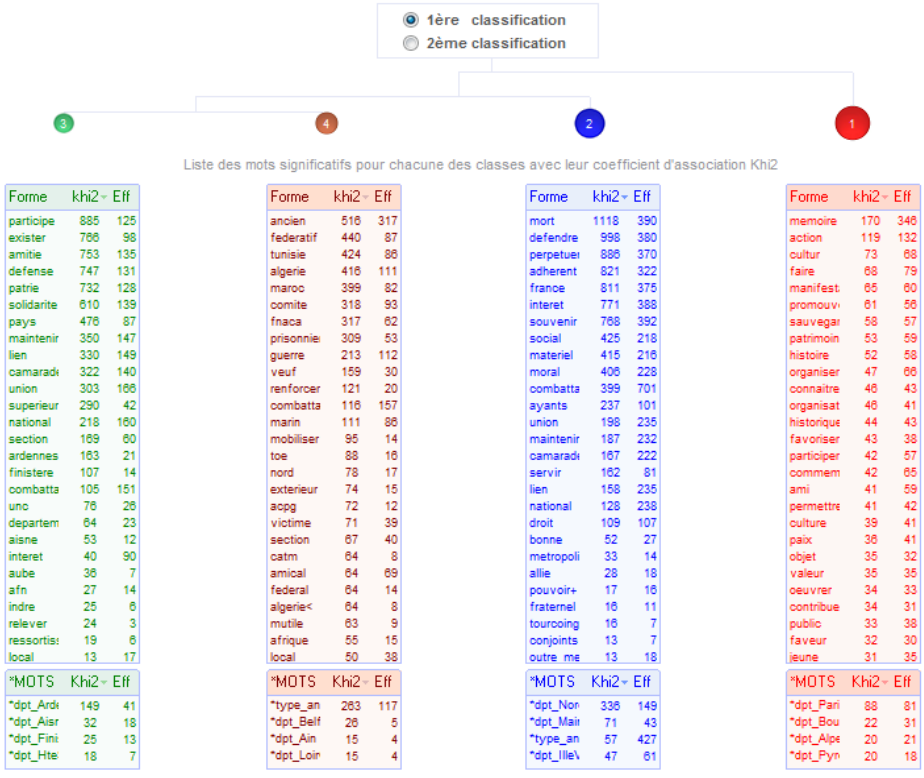


Figure 54 – Classification descendante hiérarchique de l'année 2000

Classification de l'année 2010

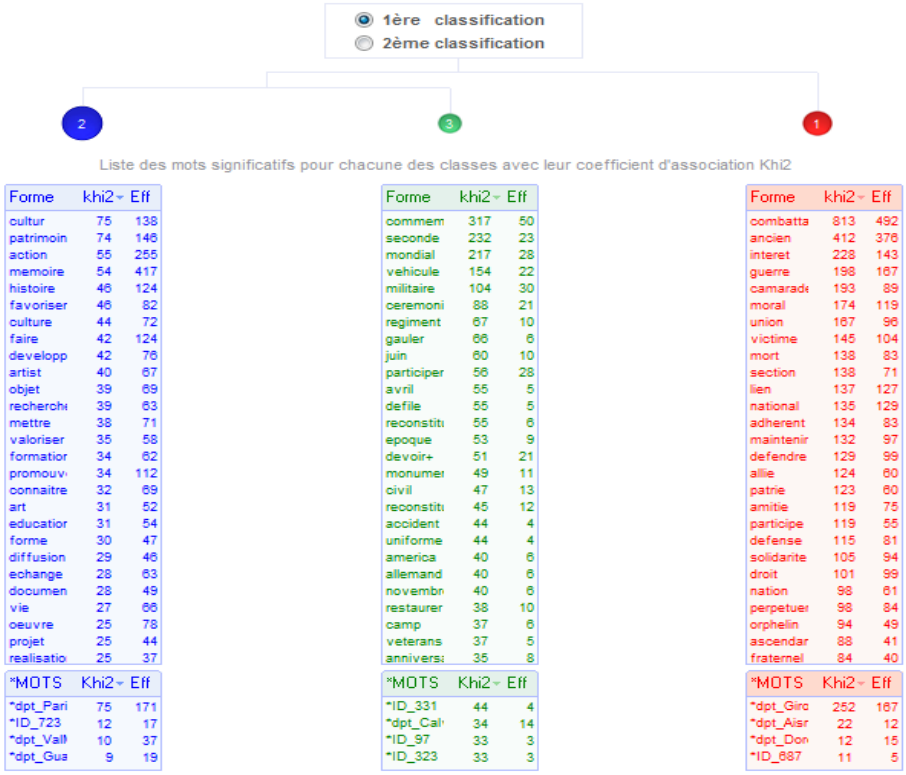


Figure 55– Classification descendante hiérarchique de l'année 2010

3.2.1 Evolution du lexique mémoriel

Quelle que soit l'année considérée, nous retrouvons la dichotomie établie lors de l'analyse du corpus dans sa globalité entre, d'un côté, une branche à dominante « anciens combattants » et de l'autre, une branche « culturelle », à laquelle appartient le terme « mémoire ». Pourtant, l'analyse disjointe des fichiers fait apparaître une différence au fil du temps dans la manière dont se constituent les classes. En effet, en 1984 et en 2000, la classe 1, à laquelle est associé le terme « mémoire », est la classe la plus homogène, celle dont les types d'énoncés sont les plus caractéristiques – et donc les plus dissemblables des autres classes. Pour ces années-là, la classe « mémoire » est ainsi la première, lors du processus de classification, à se distinguer, à se séparer des autres. Cette spécificité est associée à une couverture du corpus faible mais qui s'accroît dans le temps : en 1984, la classe 1 ne rassemble que 18% des énoncés du corpus pour atteindre 48% en 2000. Les types d'énoncés appartenant à la classe « mémoire » conservent donc leur typicité dans le temps tout en augmentant en présence.
L'organisation de ces deux années diffère ainsi de la structuration qu'a révélée l'analyse globale, dans laquelle la classe « anciens combattants » était à la fois la plus spécifique et la plus importante du corpus.

Se produit en 2010 un changement dans l'organisation des classes, qui traduit un renversement entre celle des « anciens combattants » et celle liée à la « mémoire » : à cette date, les trois classes qui traitaient des conflits passés en 1984 (anciens combattants, Afrique du Nord, souvenir) et en 2000 (anciens combattants, souvenir, solidarité) n'en forment plus qu'une seule (au poids nettement inférieur[222]), tandis que le vocabulaire de la classe « mémoire » se répartit, lui, en deux classes.

[222] En 1984, l'ensemble des classes en lien aux conflits passés représente 82% des énoncés du corpus (la classe « anciens combattants » pesant à elle seule 40%) ; en 2000, les trois classes couvrent 52% du corpus (la classe « anciens combattants » ayant drastiquement diminué à 12%) et en 2010, la classe unique « anciens combattants » représente 32% du corpus analysé.

Nous assistons donc, à la fois, à une perte de spécificité de la classe « mémoire » au profit de celle des « anciens combattants » (qui devient la première à se distinguer en 2010) et à une augmentation parallèle de sa couverture du corpus (qui devient majoritaire en 2010, avec 59% des énoncés du corpus analysé).
En comparant cette permutation temporelle entre les classes et l'organisation au niveau du corpus global, nous constatons que, si la classe « anciens combattants » demeure, du point de vue général, la plus typique et la plus importante du corpus, elle tend à perdre ces caractéristiques au fil du temps. Parallèlement, la classe « mémoire » voit ses liens à d'autres thématiques s'accroître, pour former deux classes différentes – ce qui correspond à sa représentation dans le corpus général.

Nous allons désormais nous pencher sur le vocabulaire constitutif des classes et tenter de repérer les éventuels glissements lexicaux entre elles que l'analyse statique globale ne permettait pas de montrer.

La prise en compte de la dimension chronologique révèle que ce n'est pas tant le vocabulaire associé à la classe que nous nommons « mémoire » qui varie (même si le terme lui-même perd, en 2010, la représentativité qui le caractérisait en 1984 et 2000, au profit des termes « culture » et « patrimoine ») que la scission, en 2010, d'une partie de cet univers thématique. Pris dans leur globalité, les énoncés mémoriels demeurent en effet plus ou moins les mêmes d'une année à l'autre, composés de termes relevant essentiellement du domaine culturel et historique : « mémoire », « création », « culture », « patrimoine », « histoire », etc.
Par contre, un type de discours se voit dissocié de ce groupe, en 2010, pour constituer une classe à part entière : l'univers de discours en lien à la « commémoration ». Présente en 1984 et 2000 au sein de la classe « mémoire », la commémoration caractérisait surtout la célébration d'événements historiques éclectiques ([bi-]centenaires de la révolution française, de la fondation d'une ville, de la mise en place d'un service de transport, de la construction d'une abbaye, etc.), et dont la dimension événementielle et culturelle (organisation de cérémonies, manifestations, bourses d'échanges, etc.) l'emportait sur la dimension militaire. En 2010, l'aspect commémoratif, désormais détaché de la classe « mémoire » (mais qui lui est toujours associé, en tant que subdivision de la branche « culture »), voit son lien au domaine des conflits s'autonomiser et se préciser, son vocabulaire se spécifier : les syntagmes nominaux de « seconde guerre mondiale », de « véhicule militaires », de « régiment d'infanterie », de « reconstitutions historiques » de périodes troublées, d'« uniformes et d'équipements militaires », etc. prennent le pas sur la dimension culturelle et patrimoniale qui y était auparavant attachée. Cette classe constitue ainsi une sorte de pont thématique entre le pôle strictement mémoriel et celui des anciens combattants.
A noter que, de son côté, le terme « souvenir » est comparativement toujours plus associé, quelle que soit l'année, au vocabulaire de la branche « conflits et anciens combattants ».

Ces associations de mots-clés mémoriels correspondent à celles présentées dans le corpus général ainsi qu'aux résultats de Calliope en début de période – mais non à ceux de fin de période. Mais, si l'on regarde l'évolution temporelle de la répartition de ces termes entre toutes les classes, on s'aperçoit que le terme « souvenir » voit sa spécificité se réduire au fil du temps (de classe à part en 1984 et 2000, il relève de celle des anciens combattants en 2010) et ses liens à la classe « mémoire » s'accroître en parallèle (7% des occurrences de « souvenir » relèvent de cette classe en 1984 pour atteindre 30% en en 2010, proches des 40% classés dans la catégorie « anciens combattants »). De même, le terme « commémoration », situé du côté de la branche « mémoire », voit néanmoins ses liens aux anciens combattants augmenter (aucune occurrence n'apparaît dans cette classe en 1984 alors que 21% y figurent en 2010). Le rapprochement de ces deux termes provient, en fait, de l'absence de dissociation en 2010 entre les classes « anciens combattants » et « souvenir », dont « commémoration » est proche.

3.2.2 Expressions mémorielles

Distribution des termes dans les classes[223]						
	Classe1	Classe2	Classe3	Classe4	Classe5	**Nb total occurrences**
Devoir+	19 14,84%	2 1,56%	36 28,13%	20 15,63%	18 14,06%	**128**
Lieu+	2 17%	0 11,16%	11 17,41%	36 2,23%	12 25%	**487**

L'expression « devoir de mémoire », qui apparaît dans notre corpus en 2000[224] au sein de la classe « mémoire », voit ses liens migrer du registre strictement mémoriel (2000) vers le registre guerrier et militaire (2010)[225]. En 2000, le contexte d'apparition de cette expression, même s'il comporte de nombreux liens aux conflits passés, est multiple (patrimoine, histoire, armée et anciens combattants, lien de parenté, tradition patriotique, guerre d'Espagne, CRS, etc.), parfois même l'objet de ce devoir n'est pas mentionné – comme le montre l'extrait ci-dessous :

Concordancier de **devoir+** dans le corpus

: manifestations festives et culturelles, le devoir de memoire par-rapport a ce patrimoine.
:s ou toutes actions mettant en valeur notre devoir de memoire envers les generations futures.
.a memoire historique et contribuer ainsi au devoir de memoire.
constamment le devoir de memoire; conforter le lien armee_nation; ceder, a
:e tout en creant ou en fortifiant en eux le devoir de memoire vis_a_vis du pays d' origine des parents
collectif du devoir de memoire a larodde observation et defense de tout
.aintenir notre reconnaissance; participer au devoir de memoire et aux ceremonies patriotiques; contribue
valeurs et traditions marines; perpetuer le devoir de memoire; resserrer les liens d' amitie entre anci
participer au devoir de memoire; participer aux activites de la fondation
)ir la solidarite entre tous ses membres, le devoir de memoire, le maintien des traditions patriotiques
perpetuer le devoir de memoire en organisant et assistant aux ceremonies
perenniser le devoir de memoire particulierement aupres des jeunes genera
.voriser et developper le travail relatif du devoir de memoire concernant l' exil republicain espagnol e
: siassia en tant-que resolution du probleme devoir de memoire:
: les plus eprouves et isoles; entretenir le devoir de memoire du corps des crs, de la crs 60 en-particu
satisfaire au devoir de memoire de ce-que fut notre action durant l' occu
.' accueil; faire toute action permettant le devoir de memoire; creer un lien et une solidarite entre le
participer au maintien du devoir de memoire et au droit du souvenir.
:mbres anciens combattants, et participer au devoir de memoire.
: perpetuer le souvenir de leur action et le devoir de memoire auquel ils sont tres attaches.

En 2010, nous constatons que l'univers environnant le devoir de mémoire accentue son lien aux conflits et à leur histoire, comme nous pouvons le voir ci-dessous :

[223] Comme le tableau présenté plus haut, le premier chiffre (en noir) indique la fréquence du terme dans la classe, le second chiffre (en bleu) indique le rapport entre sa fréquence et le nombre total d'occurrences du terme au sein du corpus (en %).
[224] Il est possible, et même probable, que cette expression apparaisse antérieurement mais sans que nous puissions en dater précisément l'avènement, notre corpus ne comprenant pas de déclarations entre les années 1984 et 2000.
[225] Ainsi, en 2000, 91% des occurrences de « devoir de mémoire » relèvent de la classe « mémoire » alors qu'en 2010, ce ne sont plus que 32% et 16,5% apparaissent désormais au sein de la classe « anciens combattants ». A noter que, bien que moins fréquente au sein de la classe « commémoration » (23%) que dans celle de « mémoire », elle en constitue pourtant d'après Alceste une expression typique puisqu'il l'attribue, dans sa classification, à cette classe.

Concordancier de devoir+ dans le corpus

militaires, camps de reconstitutions wwii et devoir de memoire.
du cantion de richelais conserver le devoir de memoire celebrer les commemoratives du 19 mars, accorc
le de toulouse du 10 avril 1814 perenniser le devoir de memoire aupres de la population; organiser les ceremor
d m d day 44, devoir memoire d day 44, organiser des evenements commemoratifs
pratique du devoir de memoire par l' organisation de ceremonies commemorativ
4 reconstituer, conserver le patrimoine et le devoir de memoire sur la liberation et la resistance dans la val
tes ceremonies commemoratives; transmettre le devoir de memoire aux jeunes generations; veiller a l' entretier
maquisards reconstitutions promouvoir le devoir de memoire; regrouper les personnes passionnees par l' hi
association patrimoine et devoir de memoire 39_45 restaurer au plus pres de la realite de
ation du patrimoine historique et militaire a devoir de memoire, de la poche de saint nazaire, a-travers des e
stahl_m1 mettre en valeur le devoir de memoire, les actes historiques datant de la seconde gu
participer au devoir de memoire et au respect de l' autorite de la france.
ilhan sur garonne la defense des droits et le devoir de memoire.
battant de saint denis les sens participer au devoir de memoire pour morts pour france et rassembler dionysier
le des anciens combattants unc et des amis du devoir de memoire de colleville_montgomery creer et maintenir er
assurer le devoir de memoire specifique a l' infanterie et a la perpetuatic

L'appartenance de cette expression à la classe « commémoration », en 2010, correspond à celle qu'elle occupe dans le corpus global : cela signifie que lorsque nous considérons le corpus dans sa globalité, en tant que structure en soi, le devoir de mémoire caractérise davantage cette classe, à la forte dimension symbolique, que la classe « mémoire », dont la dimension patrimoniale est plus affirmée.
L'expression « lieu de mémoire » caractérise, quant à elle, davantage le domaine patrimonial, comme c'est le cas en 2000 (et dans le corpus global) :

presbytere de remoray restaure en lieu de memoire maison du patrimoine, destine a la c
r du musee pour qu' il devienne un lieu de memoire et un centre culturel et civique de
ere_port; retrouver la memoire des lieux; perenniser le souvenir et les savoir_faire de
belleville preserver la memoire du lieu et la maintenir la plus vivante possible; defen
possible; defendre et organiser le lieu pour le rendre plus vivant, plus democratique,
rences, communications, visite des lieux de vie, publications, reeditions des oeuvres,
nation dudit prieure et des autres lieux de memoire;
urs et touristes l' histoire de ce lieu.
lieux de memoire communs franco_quebecois ldmcfq inv

Ou en 2010 :

l' ecole militaire lieu de memoire contribuer a restituer et met
n environnement culturel et de ses lieux de memoire.
uvre de projets visant a animer ce lieu de memoire par des actions culturelles t
histoire, de ses quartiers de ses lieux connus et moins connus et de ses habita
couvrir et/ ou se reapproprier les lieux de memoire.
ce et a l' etranger, dans tous les lieux publics ou prives qui s' y pretent, ain
u village et son utilisation comme lieu de reunion et de rencontre conviviale;
e collectif pour la creation d' un lieu de memoire industrielle et ouvriere du p

L'expression « lieu de mémoire » est beaucoup moins figée que celle de « devoir de mémoire » : le terme « lieu » apparaît en effet dans des contextes plus variés que ceux de « devoir de mémoire » (mémoire du lieu, organiser le lieu, visite des lieux, histoire de ce lieu, lieu de réunion, etc.).Est-ce à dire que l'usage de cette expression est moins stabilisé, qu'elle est moins « socialement signifiante » que celle de « devoir de mémoire » ? Ou que notre corpus fait état d'une présence plus importante de cette dernière ?

3.2.3 Thématiques « communautaristes »

De même qu'avec Calliope, nous avons voulu savoir comment se présentaient et évoluaient, sous Alceste, les termes relevant de thématiques dites « communautaristes ».
Précisons, tout d'abord, qu'en raison de ses principes de fonctionnement, un certain nombre de termes ne sont pas considérés comme typiques par Alceste. Ces termes (esclave, musulman, identité, immigration, migration, Shoah, DOM TOM, breton, Bretagne) sont en effet trop peu fréquents pour pouvoir être davantage associé à l'une des classes élaborées, et ce, même en se situant au niveau général du corpus unique : leur très faible fréquence empêche qu'ils puissent cooccurrer suffisamment avec d'autres termes pour pouvoir relever de fragments typiques d'une

classe. La seule façon d'étudier ces termes consiste donc à en analyser la distribution – ce que nous n'avons pas eu le temps de faire dans le cadre de ce travail.
D'autres termes (racisme, antisémitisme, juif, banlieue) ne sont associés à une classe qu'au niveau du corpus global mais non à celui des sous-corpus annuels.

Distribution des termes dans les classes						
	Classe1	Classe2	Classe3	Classe4	Classe5	**Nb total occurrences**
rac+	0	4 14,29%	19 67,86%	1 3,57%	0	**28**
antisemitis+	0	0	8 80%	0	0	**10**
jui+f	0	0	8 25,81%	12 38,71%	0	**31**
banlieue+	1 14,29%	0	0	0	3 42,86%	**7**
revendicati+	10 83,33%	0	2 16,67%	0	0	**12**

Racisme
D'un point de vue global, les termes « racisme » et « antisémitisme » relèvent davantage de la classe « commémoration » (classe qui regroupe les luttes contre toutes les formes de discriminations), cooccurrant d'ailleurs souvent ensemble. Cette caractéristique est confirmée par l'analyse de leur distribution dans les classes, qui est très inégale (fortement concentrée au sein de la classe 3).

Concordancier de **racisme** dans le corpus

es victimes individuelles ou collectives du **racisme** et de l' antisemitisme; lutter contre toute
tionales d' education et de lutte contre le **racisme** et le harcelement moral, de l' antisemitism
le negationnisme, la xenophobie et tous les **racismes**, elle se prononce pour la securite des per
d' inspiration fasciste a la xenophobie, au **racisme** et a l' antisemitisme.
h dans les landes, combattre les ideologies **racistes**, xenophobes, antisemites et lutter pour la
sion centrale de l' enfance, cce, contre le **racisme** et l' antisemitisme, pour un judaisme progr
emes buts, sans distinction d' opinions, de **race** ou de religion.
ure ethnique, excluant toute distinction de **race**, de nationalite, d' opinions politiques ou rel
gies totalitaires, fascistes, xenophobes et **racistes**.
combattants sans distinction d' opinion, de **race** ou de religion.
attre par tous les moyens en son pouvoir le **racisme** et l' antisemitisme; promouvoir les droits
emes buts, sans distinction d' opinions, de **race** ou de religions.
ibertes; lutter contre toutes les formes de **racisme**, le fascisme et le totalitarisme;
ligue internationale contre le **racisme** et l' antisemitisme d' aix et des pays d' a
attre par tous les moyens en son pouvoir le **racisme** et l' antisemitisme;
position a toutes formes d' exclusion et de **racisme**.

Concordancier de **antisemitisme** dans le corpus

iduelles ou collectives du racisme et de l' **antisemitisme**; lutter contre toute forme de discriminatior
le racisme et le harcelement moral, de l' **antisemitisme**, les discriminations et les guettos; poursui
sciste a la xenophobie, au racisme et a l' **antisemitisme**.
e l' enfance, cce, contre le racisme et l' **antisemitisme**, pour un judaisme progressiste et les valeu
les moyens en son pouvoir le racisme et l' **antisemitisme**; promouvoir les droits de la personne humair
gue internationale contre le racisme et l' **antisemitisme** d' aix et des pays d' aix, licra, aix et pa
les moyens en son pouvoir le racisme et l' **antisemitisme**;
gue internationale contre le racisme et l' **antisemitisme**, licra, federation de la cote d' azur sectic
gue internationale contre le racisme et l' **antisemitisme**, section de nice, licra nice, reprenant l' c
iduelles ou collective du racisme et de l' **antisemitisme**;

A noter que 7 occurrences sur 28 de la racine « rac+ » correspondent au terme « race » et non à « racisme » (en raison du processus de lemmatisation), et ce sont précisément les occurrences qui apparaissent dans les classes autres que la classe 3. Elles correspondent notamment à des déclarations d'anciens combattants, au sein de la classe 2 (« association ouverte à tous ceux qui ont servi sous les armes, sans distinction d'opinion, de race ou de religion »).

Juif

En ce qui concerne le terme « juif » (et ses dérivés, juive(s) et juifs), nous observons une certaine évolution entre 2000 et 2010[226], qui se traduit par une diversification des thèmes qui lui sont associés. Ainsi, en 2000, le terme « juif » cooccurre fréquemment avec celui de « mémoire » : il s'agit surtout d'établir la mémoire des enfants juifs morts en déportation (rechercher le nom des enfants et en rappeler l'existence).

En 2010, le contexte d'apparition du terme « juif » s'élargit, pour intégrer une dimension patrimoniale qui lui faisait défaut en 2000. Cette dimension patrimoniale et culturelle est celle qui prédomine lorsque l'on interroge le corpus au niveau global (la classe 4 rassemble près de 40% des occurrences de ce terme).

En 2000 :

Concordancier de **juif** dans le corpus

association pour la memoire des enfants **juifs** deportes du 5e, amejd 5e, rechercher des noms d
amejd 5e, rechercher des noms des enfants **juifs** scolarises sous l' occupation allemande dans le
association pour la memoire des enfants **juifs** d' ozoir, amejo, assurer la commemoration des e
mejo, assurer la commemoration des enfants **juifs** d' ozoir_la_ferriere morts en deportation et de
onde guerre mondiale ont aide ou sauve des **juifs**.
association pour la memoire des enfants **juifs** deportes du 18e, amejd, rappeler la memoire des
8e, amejd, rappeler la memoire des enfants **juifs** deportes du 18e arrondissement de paris, pendan
association pour la memoire des enfants **juifs** deportes du 2e, amejd 2e, rappeler la memoire d
amejd 2e, rappeler la memoire des enfants **juifs** deportes du 2e arrondissement pendant la second
oujda_beth_chalom, association des **juifs** originaires d' oujda et de sa region entretenir
moire, le patrimoine et les traditions des **juifs** d' oujda et de sa region.

En 2010 :

Concordancier de **juif** dans le corpus

patrimoine et cultures des **juifs** du liban pcjl preserver la memoire de la communaute juive
an pcjl preserver la memoire de la communaute **juive** du liban, notamment par la collecte de documents, de text
re, memoire, culture, art, patrimoine, pensee **juive**, commentaires religieux.
ments relatifs a l' histoire de la resistance **juive**;
t educatifs du consistoire et des communautes **juives** de france;
ucjf, union des communautes **juives** de france contribuer a la construction et a la conservat
imoine culturel et artistique des communautes **juives** de france;
evoquer accueil des refugies et sauvetage des **juifs** sur plateau vivarais_lignon entre 1939 et 1944.
numents funeraires dependants d' associations **juives** en desherence;
ciations loi 1901 et des societes mutualistes **juives** en difficulte; sauvegarde et entretien des caveaux colle
oution des places disponibles a des personnes **juives** en isolees ou dont les familles sont en difficulte.
paradis et de l' espace memoire des residents **juifs** de la moi.
e de la resistance et du genocide des enfants juifs dans les landes perpetuer, promouvoir et defendre les va
ion r m k soutien actif pour toutes activites **juives** dans-le-domaine-de l' education de la perennite du souve

Au niveau global :

[226] Même si le terme est toujours principalement associé à la même classe, celle de « mémoire », en raison de sa plus forte cooccurrence avec d'autres termes au sein d'énoncés de cette classe.

Concordancier de **juif** dans le corpus

	patrimoine et cultures des	**juifs** du liban pcjl preserver la memoir
preserver la memoire de la communaute	**juive**	du liban, notamment par la collec
oire, culture, art, patrimoine, pensee	**juive**,	commentaires religieux.
les programmes audiovisuels d interet	**juif**	ainsi que susciter et encourager l
s institut de la memoire audiovisuelle	**juive**	i m a j.
r et developper une memoire collective	**juive**	et promouvoir les aspects artisti
et de l' espace memoire des residents	**juifs**	de la moi.
s loi 1901 et des societes mutualistes	**juives**	en difficulte; sauvegarde et ent
des places disponibles a des personnes	**juives**	en isolees ou dont les familles
resistance et du genocide des enfants	**juifs**	dans les landes perpetuer, promou
elatifs a l' histoire de la resistance	**juive**;	
accueil des refugies et sauvetage des	**juifs**	sur plateau vivarais_lignon entre
oujda_beth_chalom, association des	**juifs**	originaires d' oujda et de sa reg
e, le patrimoine et les traditions des	**juifs**	d' oujda et de sa region.
tifs du consistoire et des communautes	**juives**	de france;
ssociation pour la memoire des enfants	**juifs**	d' ozoir, amejo, assurer la comme
, assurer la commemoration des enfants	**juifs**	d' ozoir_la_ferriere morts en dep
ssociation pour la memoire des enfants	**juifs**	deportes du 2e, amejd 2e, rappele
jd 2e, rappeler la memoire des enfants	**juifs**	deportes du 2e arrondissement pen
guerre mondiale ont aide ou sauve des	**juifs**.	
k soutien actif pour toutes activites	**juives**	dans-le-domaine-de l' education
ucjf, union des communautes	**juives**	de france contribuer a la constr
culturel et artistique des communautes	**juives**	de france;

A noter au niveau global que le terme « juif » apparaît également dans des énoncés (ici, en vert) considéré comme typique de la classe 3 (lutte contre les discriminations), qui rassemble près de 26 de ses occurrences.

Banlieue

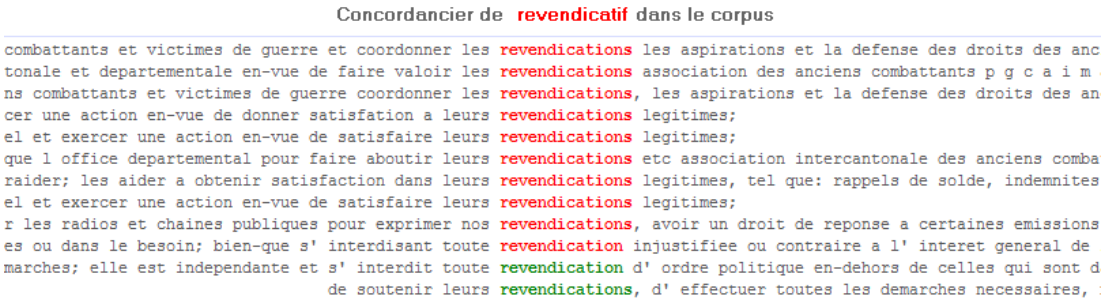
Le terme de « banlieue », trop peu fréquent (7 occurrences totales) pour être caractérisé dans les sous-corpus annuels, est davantage associé, quant à lui, à la classe « échanges » (en rose), en lien avec l'organisation de manifestations ou le développement d'études sur ce que cette notion représente. Lorsqu'il apparaît au sein de fragments de la classe 1, c'est pour désigner le lieu d'exercice de l'association déclarante.

Concordancier de **banlieue** dans le corpus

.a question des representations des villes de la	**banlieue**	et de celles de la ville centre,
	banlieues	creation d une manifestation internationale a car
:isienne et qui aurait pour titre a festival des	**banlieues**	du monde le descriptif du projet figurant a l art
)mbattants et victimes de guerre section orleans	**banlieue**	fleury_les_aubrais, saran association republicaine
)mbattants et victimes de guerre section orleans	**banlieue**	fleury_les_aubrais, saran.
objet reflexion et recherche sur le concept	**banlieue**	dans le monde etude pratique de ce phenomene de ci
unite a tout patrimoine commun aux villes de la	banlieue,	a leurs habitants et a l' histoire socio urbaine

Revendication

A noter qu'un terme « revendicatif », qui n'apparaît pas dans les résultats de Calliope, est caractérisé comme relativement plus typique de la classe « anciens combattants » (83% des occurrences de ce terme apparaissent dans cette classe) dans Alceste (au sein du corpus global) : il s'agit du mot « revendication » lui-même. Ses autres apparitions concernent la classe 3 (lutte contre les discriminations) :



3.2.4 Evolution temporelle de la répartition par département et par type d'annonce

Après avoir montré les glissements thématiques et lexicaux que révélait une analyse chronologique du corpus, nous désormais aborder l'évolution temporelle de la répartition géographique des classes et celle par type d'annonce.

Pour en simplifier la lecture, nous avons synthétisé les informations sur la distribution géographique des classes (et des types d'annonces) dans le tableau ci-dessous :

	Département	**Type d'annonce**
1984		
Classe 1 **Mémoire / Commémoration**	Paris (93) Pyrénées (9) Aude, Gironde, Pas-de-Calais (- 3)	Création (12) Modification (- 12)
Classe 2 **Combattants d'Afrique du Nord**	Isère (26) Ardennes, Haute-Loire (15) Paris (- 5)	Création (0) Modification (-0)
Classe 3 **Souvenir**	Pyrénées (11) Nord, Paris (- 5)	Création (8) Modification (2)
Classe 4 **Anciens combattants**	Nord (15) Aude (14) Paris (- 13)	Modification (2) Création (- 34)
2000		
Classe 1 **Mémoire**	Paris (88) Bouches-du-Rhône (22) Nord (- 119) Ardennes (- 66)	Création (8) Modification (- 8)
Classe 2 **Souvenir**	Nord (336) Maine-et-Loire (71) Ille-et-Vilaine (47) Paris (- 35)	Création (57) Modification (- 57)
Classe 3	Ardennes (149)	Création (0)

Solidarité	Aisne (32) Nord (- 22)	 Modification (0)
Classe 4 **Anciens combattants**	Territoire de Belfort (26) Ain, Loir-et-Cher, Charente, Lot-et-Garonne (15) Nord (- 23)	Modification (263) Création (- 268)
2010		
Classe 1 **Anciens combattants / souvenir**	Gironde (252) Aisne (22) Paris (- 68)	Création (0) Modification (0)
Classe 2 **Mémoire**	Paris (75) Val-de-Marne (10) Calvados (- 24)	Modification (4) Création (- 2)
Classe 3 **Commémoration**	Calvados (34) Val-d'Oise (23) Gironde (- 19)	Création (6) Modification (- 7)

Répartition géographique

Nous constatons, de même que lors de l'analyse globale, que le thème de la « mémoire » est toujours davantage lié, sur les trois années, à Paris. La dimension culturelle propre à ce département-ville en constitue certainement une des raisons. A noter, qu'en 2000, que la probabilité relative, pour la classe « mémoire », de ne pas être associée au département du Nord est supérieure à celle d'apparaître dans la capitale. Cette année-là, un lien extrêmement fort se dessine par contre entre ce même département du Nord et la classe « Souvenir ». Ce lien est si important qu'il apparaît au niveau global – comme nous l'avons vu plus haut.

Après ne pas avoir été marquée par un ancrage territorial précis en 1984 et en 2000, la classe des anciens combattants est plus fortement associée, en 2010, au département de la Gironde.

Vocabulaire représentatif de Paris (2000)

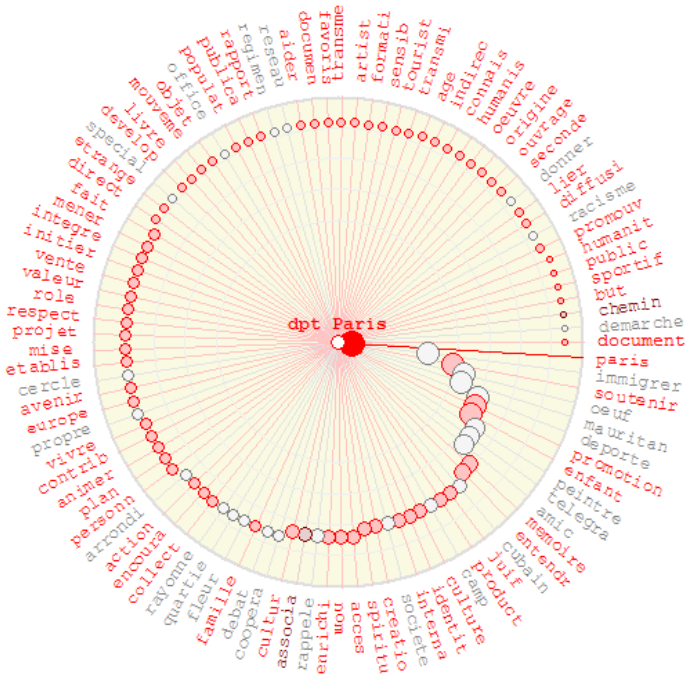


Figure 56 – Disque Département Paris / Formes réduites

Vocabulaire représentatif de la Gironde (2000)

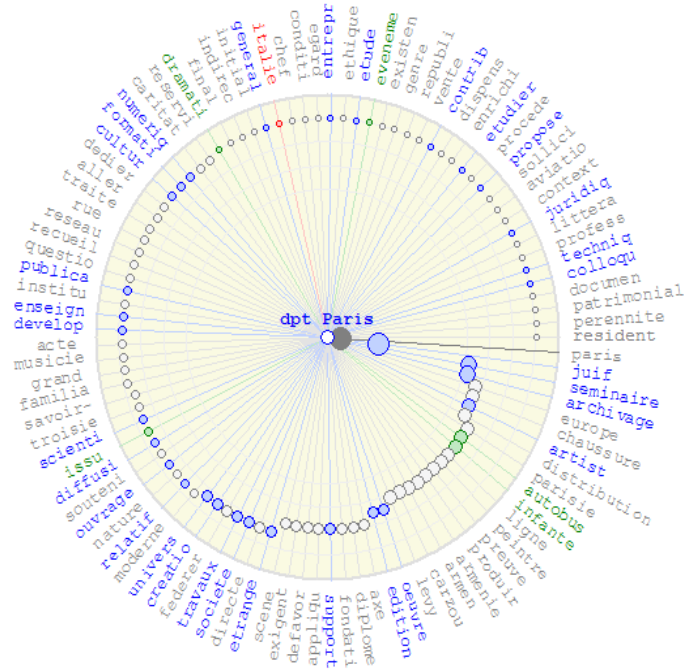


Figure 57 – Disque Département Gironde / Formes réduites

Distribution par type d'annonce

En ce qui concerne le type d'annonce, les modifications d'associations caractérisent essentiellement les anciens combattants, et les créations la classe « souvenir ». Ces caractéristiques se retrouvent dans le corpus général. Les autres classes sont peu affectées par cette variable.

Pour conclure, il faut ajouter que l'ensemble des éléments présentés ici demandent, pour pouvoir être correctement appréhendés, à être resitués dans leur contexte d'apparition, notamment politique, institutionnel, social et culturel. Il faudrait également, par un retour aux textes, voir, par exemple, quel type d'association est modifié ou créé. Lors d'une modification, constate-t-on une modification du vocabulaire mémoriel ? Dans quel sens ?

4 Comparaisons des résultats obtenus avec Calliope et Alceste

Nous allons, dans cette partie, procéder à une analyse comparative des résultats[227] issus des deux logiciels de lexicométrie utilisés dans le cadre de ce travail, Calliope et Alceste. Bien que ces logiciels se fondent tous les deux sur des méthodes de statistiques textuelles (construction de tableaux lexicaux, classification hiérarchique), les divergences dans l'implémentation de ces méthodes et dans la définition de l'unité textuelle de base génèrent des différences au niveau des résultats. Nous présenterons donc ces divergences, en les reliant, autant que possible, à la méthode sous-jacente qui a présidé à leur élaboration. Nous tenterons de montrer, au fur et à mesure, les complémentarités qui se dégagent de cette comparaison.

4.1 Similarités : des résultats globalement concordants

Cependant, avant d'aborder l'analyse de la complémentarité de ces outils, nous allons auparavant en présenter les points communs, en termes de résultats.
Si nous nous situons à un niveau très général, nous constatons que les résultats obtenus via le traitement du corpus par les deux logiciels, Calliope et Alceste, sont globalement concordants et que des dichotomies thématiques similaires sont révélées entre les déclarations. Nous observons en effet une même distinction entre deux univers séparés, le monde des anciens combattants d'un côté et celui de la mémoire, liée au domaine culturel, de l'autre. Cette similarité de résultats s'explique bien sûr par l'unicité du corpus traité, le vocabulaire attaché aux déclarations d'associations étant le même.
Ces deux domaines distincts apparaissent aussi bien dans les diagrammes stratégiques et les clusters élaborés par Calliope que dans les classes de fragments de discours construites par Alceste.

Rappelons, pour mémoire, la constitution des diagrammes stratégiques sur les 3 années :

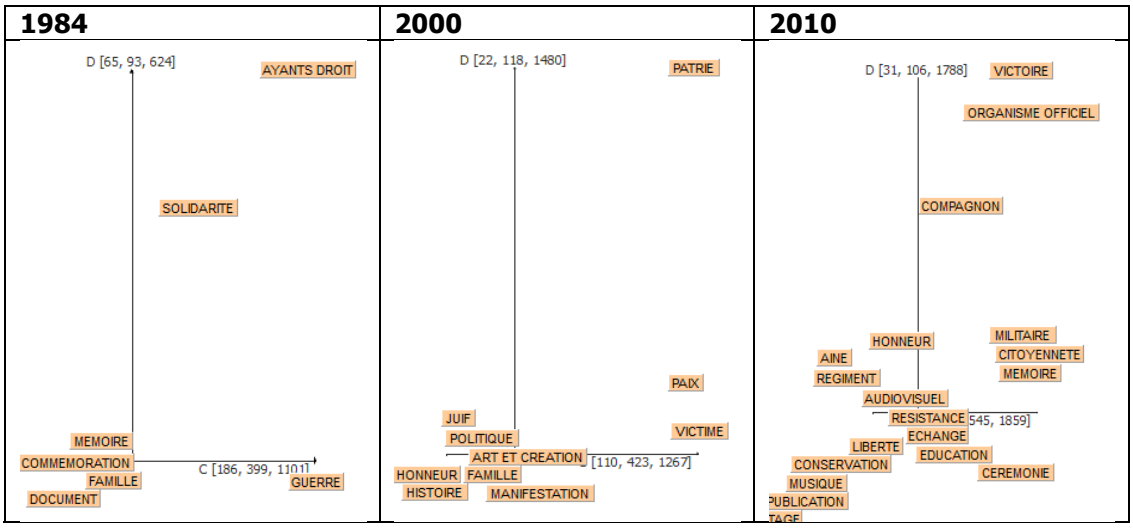


Figure 58 – Diagrammes stratégiques (1984, 2000, 2010)

[227] Nous n'abordons pas ici la comparaison entre les principes de fonctionnement de ces deux outils et les types de (pré-)traitements requis, qui ont déjà été présentés dans la première et la deuxième parties de ce mémoire.

Les thématiques des deux domaines y figurent clairement : d'un côté, « ayants-droits », « solidarité », « guerre » en 1984, « patrie » », « paix », « victimes » en 2000 ou « victoire », « organisme officiel » ou encore « militaire », pour n'en citer que quelques-uns, appartiennent à la thématiques des anciens combattants ; de l'autre, du côté de la mémoire, nous trouvons les termes « mémoire », « document » en 1984, « art et création », « histoire », « manifestation » en 2000 et « mémoire », « audiovisuel », « éducation » en 2010.
Cette répartition thématique globale se retrouve dans les résultats produits par Alceste, dont nous ne reproduisons ici que la classification établie sur le corpus général :

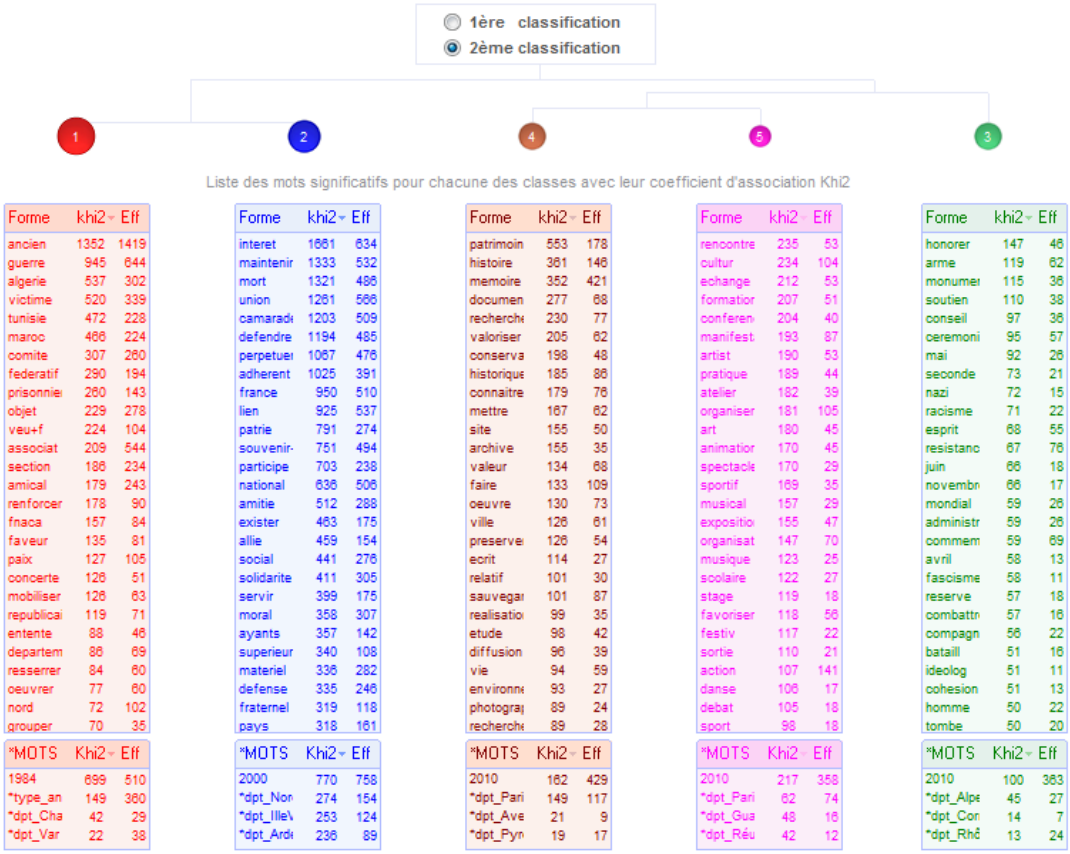


Figure 59 – Classification descendante hiérarchique du corpus global

Autre similarité constatée entre les deux outils, la tendance inverse qui affecte les domaines des anciens combattants et de la mémoire. D'un côté, se produit une diversification des thèmes associés à la « mémoire », ainsi que la baisse du poids des anciens combattants. Nous avons vu avec Calliope que le cluster « Mémoire » devenait de plus en plus important, au cours des trois années de notre corpus, sa fréquence et ses liens (internes et externes) à d'autres termes augmentant dans le temps. De même, Alceste traduit bien cette augmentation de l'ampleur thématique rattachée à la classe « Mémoire » dont le poids s'accroît et les thématiques se subdivisent en deux classes. Pour les deux logiciels, en 1984 et 2000, les thèmes associés aux anciens combattants sont les plus importants, tandis que l'année 2010 révèle une forte croissance de la thématique mémorielle, *stricto sensu*.

4.2 Spécificités des deux outils

Mais, une fois énoncées ces similarités générales entre résultats, une analyse plus fine des deux démarches fait état de différences notables qui tiennent, répétons-le, aux principes

guidant l'élaboration des résultats. Car, derrière cette homologie générale entre résultats, des spécificités apparaissent, qui résultent de vues dissemblables sur un même corpus. Fondés sur des méthodes statistiques différentes, la façon qu'ont ces deux outils de définir, compter et classer les éléments divergent nécessairement. Les résultats produits en sont la démonstration, dont notre analyse même en porte la trace : le type de traitements et les différentes représentations graphiques des résultats orientent l'analyse qui en est faite. Ainsi, l'accent est mis, dans Calliope, sur les termes et les liens qu'ils entretiennent alors qu'avec Alceste, nous observons un dévoilement d'univers de discours distincts, se focalisant sur le contexte d'apparition des termes, leur environnement linguistique proche. Nous essaierons de montrer en quoi ces différences constituent une complémentarité pour l'analyse du corpus.
Pour ce faire, nous nous baserons sur la grille comparative que nous avons établie, en en développant chaque point.

	Calliope	**Alceste**
Spécificités des outils > complémentarité des démarches		
Choix des « unités textuelles représentatives » et résultats produits	Représentation du **lexique** du corpus > accent mis sur les **réseaux de mots et leurs cooccurrences** > validation humaine du lexique d'annotation **Point fort** : Résumé thématique d'un corpus (diagramme stratégique), rapidement accessible **Points faibles** : Risque de se perdre dans l'analyse des multiples liens (internes et externes) entre termes + Risque d'une mauvaise validation du lexique + Absence de concordancier	Représentation des **univers de discours** du corpus > accent mis sur les **contextes d'apparition** et fragments de discours (UCE, segments répétés) **Points forts** : Distinction entre fragments de discours spécifiques (analyse fine du vocabulaire, avec catégorisation grammaticale détaillée) + Stabilité de l'analyse (double classification) + Mise en contexte multiple des termes (concordancier, segments répétés) **Point faible** : Seuls les points saillants du discours sont perceptibles
Dimension temporelle de l'analyse	**Analyse chronologique dynamique** > comparaison temporelle des lexiques (Tendances) **Point fort :** Termes / thèmes stables, émergents, déclinants ou fluctuants sur une période	**Analyse instantanée « spécifique »** > mais possibilité de croiser les résultats avec variable temporelle **Point faible** : Absence de comparaison temporelle dynamique des vocabulaires

Points forts / spécificités de Calliope	Points forts / spécificités d'Alceste
> Mise en valeur du **réseau de cooccurrences** de termes d'un corpus, associée à une **analyse chronologique**, qui permet de révéler l'évolution des univers thématiques, lexicaux dominants ou à la marge, sur une période. > Représentation de thèmes mineurs dans la tour d'ivoire, de termes émergents, stables, fluctuants ou déclinants, et de voir leur évolution temporelle (analyse des tendances).	> Distinction fine entre **différents univers de discours**, qui en montre les spécificités (fragments de discours spécifiques d'une classe), associée à une **analyse détaillée des catégories grammaticales**. > Mise en contraste de la multiplicité des mondes lexicaux, par un accès facilité aux énoncés typiques et à leurs opposés.

4.2.1 Choix des « unités textuelles représentatives » : des résultats complémentaires – Exemple des mots-clés mémoriels

Pour les deux outils retenus, la définition de l'unité textuelle représentative diffère. Nous considérons ici les « unités textuelles représentatives » par opposition aux unités textuelles de base, sur lesquelles se fonde le dénombrement et la comparaison statistiques (dans les deux cas, il s'agit de mots et de leurs cooccurrences). Les unités représentatives sont, pour nous, les unités issues de l'analyse (méthode de classification ascendante ou descendante), présentées en résultats « typiques » de l'outil.
Ainsi :
- Pour Calliope, qui est un logiciel issu de la recherche documentaire fondé sur la classification ascendante hiérarchique, les documents du corpus sont analysés et représentés via leur lexique, ie les mots pleins et leurs cooccurrences (descripteurs), regroupés au sein de clusters de termes.
- Pour Alceste, qui applique une classification descendante hiérarchique, le découpage des textes du corpus se fonde également sur la distribution des mots pleins mais seulement en tant qu'ils appartiennent à des unités de contexte, ie des fragments de discours au sein desquels l'outil repère la présence / absence significative de mots.

Ces deux méthodes produisent des résultats nécessairement différents. Pour Calliope, l'analyse des termes et de la force des relations qu'ils entretiennent conduit à l'élaboration de groupes de mots (clusters), positionnés sur un diagramme stratégique. Alceste, quant à lui, propose une classification d'énoncés types, comprenant un ensemble de mots liés représentatifs d'une classe. Nous associons cette différence entre outils à la méthode sous-jacente : classification hiérarchique ascendante (chaque mot isolé constitue le point de départ de l'analyse, puis est associé à d'autres mots au sein de cluster) d'un côté, et descendante (le corpus considéré comme un tout, puis distingué en énoncés) de l'autre.

La différence que ces méthodes produisent en termes de résultats se saisira mieux au travers d'exemples – que nous prendrons parmi les mots-clés mémoriels qui ont guidé notre analyse. Nous verrons que, selon les outils, les relations entre ces mots-clés, bien que classées différemment, entretiennent néanmoins des correspondances.
Ainsi de la relation entre « anciens combattants » et « commémoration ». En 2010, une même déclaration est classée de deux manières différentes par Calliope et Alceste. Le vocabulaire de la déclaration du Comité local de Marcilloles[228] (Isère), rattaché à la Fnaca,

[228] Le texte de cette déclaration, créée en 2010, est le suivant : Titre : « Comité local de la Fnaca de Marcilloles (Fédération nationale des anciens combattants d'Algérie, Maroc et Tunisie) » et Objet : « Assurer la sauvegarde des droits matériels et moraux des **anciens combattants** en Algérie, Maroc et Tunisie ; renforcer leurs liens de camaraderie et solidarité ; coordination, impulsion et soutien de l'action du comité local ; représentation du comité local au sein de toutes les instances

est associé avec Calliope au cluster « Militaire » car plusieurs de ses termes cooccurrent, dont « anciens combattants » et « commémoration ».

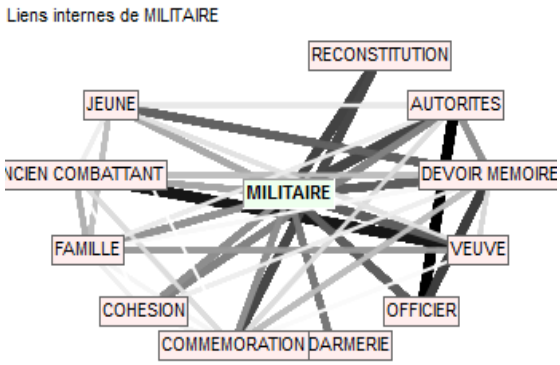


Figure 60 – Cluster « Militaire » (2010)

Dans Alceste, cette déclaration, découpée en 3 UCE, est considérée comme typique de la classe « Anciens combattants », à laquelle le terme « commémoration » n'est pas principalement associé (même si, comme nous l'avons vu plus haut, 17% de ses occurrences y figurent) :

Unité n° 921 Classe 1 Khi2 = 14 Individu n° 350 *ID_1137 *dpt_Isère *type_ann_1
comite local de la fnaca de marcilloles, federation nationale des anciens combattants d' algerie, maroc et tunisie, assurer la sauvegarde des droits materiels et moraux des anciens combattants en algerie, maroc et tunisie;
Unité n° 922 Classe 1 Khi2 = 1 Individu n° 350 *ID_1137 *dpt_Isère *type_ann_1
renforcer leurs liens de camaraderie et solidarite; coordination, impulsion et soutien de l' action du comite local; representation du comite local au sein de toutes les instances departementale, regionale et nationale du monde combattant;
Unité n° 923 Classe 1 Khi2 = 4 Individu n° 350 *ID_1137 *dpt_Isère *type_ann_1
oeuvrer en faveur de la paix, notamment par la commemoration du 19 mars 1962, cessez_le_feu en algerie.

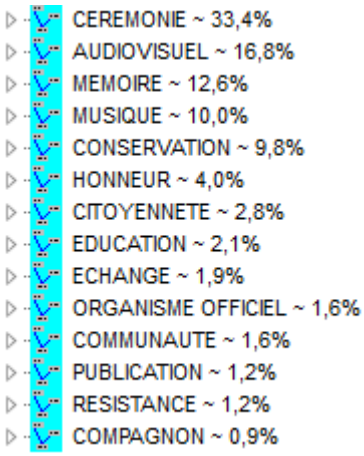
Le terme « commémoration » n'est pas associé à la classe (en gris dans l'exemple) car trop distinct de l'univers de discours propre aux anciens combattants (en rouge), ie pas assez cooccurrent avec son vocabulaire spécifique, identifié par Alceste comme se focalisant sur les liens de camaraderie et la défense des intérêts moraux et matériels. Cette occurrence du terme n'est pas non plus associé ici à la classe « Commémoration » (il n'apparaît ainsi pas en vert dans l'exemple ci-dessus, qui est la couleur de la classe « Commémoration »), dans la mesure où celle-ci est caractérisée par sa proximité thématique et lexicale à l'univers culturel et événementiel (accent mis sur les cérémonies, reconstitutions ou autres événements), comme nous pouvons le voir ci-dessous :

Unité n° 94 Classe 3 Khi2 = 40 Individu n° 38 *ID_189 *dpt_Aube *type_ann_1
comite d' organisation des fetes et ceremonie d' auxon, cofca, organiser les fetes, commemorations, ceremonies, receptions officielles sur le territoire de la commune d' auxon, notamment les 14 juillet, 1er et 11 novembre, 8 mai, 18 juin,

Cette divergence de traitements peut constituer un problème, temporaire, lors de l'interprétation des résultats. D'un côté, « anciens combattants » et « commémoration » sont fortement reliés (Calliope), de l'autre, ils sont dissociés (Alceste).
Pour dépasser cette apparente incohérence, il faut analyser les liens qu'entretiennent ces termes. Ainsi, dans Calliope, le terme « commémoration » est membre du cluster « Militaire », dans lequel figure notamment le terme « reconstitution » (aspect culturel). Il a, de plus, comme principal lien externe le terme « Cérémonie », pour un tiers de ses liens. Cela signifie qu'ils apparaissent donc très souvent ensemble.

---

départementale, régionale et nationale du monde combattant ; œuvrer en faveur de la paix, notamment par la **commémoration** du 19 mars 1962, cessez-le-feu en Algérie » [Nous soulignons]



Dans Alceste, si « commémoration » décrit une classe à part entière, située dans la branche culturelle de l'arbre hiérarchique, son lexique caractéristique renvoie au domaine militaire, dont voici un extrait :

Forme	khi2	Eff
commem	317	50
seconde	232	23
mondial	217	28
vehicule	154	22
militaire	104	30
ceremoni	88	21
regiment	67	10

Dans les deux outils d'analyse de données textuelles, le terme commémoration entretient donc un lien à la fois au domaine culturel et au domaine militaire, mais avec une mise en valeur de l'un ou de l'autre selon l'outil. La distinction entre les manières de le classer relève des méthodes sous-jacentes aux outils, comme nous l'avons vu : le terme peut être considéré soit comme une unité en soi (une forme graphique réduite) dont il s'agit d'étudier les liens aux autres mots à l'échelle du corpus, sans que son contexte d'apparition proche (le segment de phrase) en conditionne la classification ; soit comme un terme s'inscrivant dans un fragment de phrase définissant un univers de discours particulier. N'est prise en compte, dans Alceste, pour caractériser une classe relativement à une autre, que la cooccurrence de mots pleins appartenant à un même segment[229], motif qui se répète dans le corpus. Ces deux approches offrent ainsi une vision complémentaire du corpus.

Autre exemple, cette fois-ci à propos du terme « souvenir ». Les liens que ce terme entretient avec les autres mots-clés connaissent la même divergence de traitement que ceux de « anciens combattants », « commémoration » et « mémoire » que nous venons de voir. Ainsi, dans Calliope, le mot « souvenir » appartient, en 2010, au cluster « Mémoire » :

[229] Voir, à ce propos, REINERT M., « Postures énonciatives et mondes lexicaux stabilisés en analyse statistique de discours », *op. cit.*

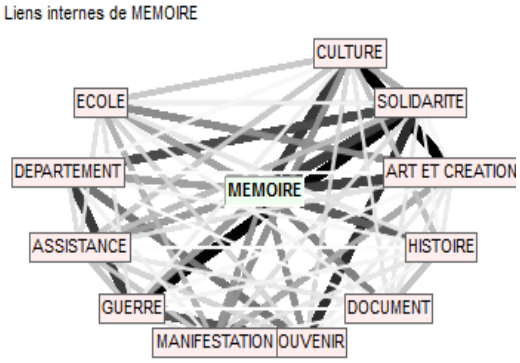


Figure 61 – Cluster « Mémoire » (2010)

Tandis que dans Alceste, il relève de la classe « Anciens combattants » (en rouge) :

Présences significatives de la classe 1

Mots analysés	Khi2	Nbre d'unités classe1	Total unités classées	Nbre d'unités en %	Marqueur grammatical
retraite	7	11	18	61%	Noms
rhin	6	3	3	100%	Lieux, pays
s	2	8	16	50%	Formes non reconnues
sacre	13	6	6	100%	Adjectifs et adverbes
sacrifice	15	13	17	76%	Noms
saugeron	8	4	4	100%	Formes non reconnues
section	138	65	66	98%	Noms
servir	81	53	63	84%	Verbes
ses	22	112	248	45%	Marqueurs de la personne (mots outils)
social	20	74	153	48%	Adjectifs et adverbes
soldat	10	9	12	75%	Noms
solennel	8	4	4	100%	Formes reconnues mais non codées
solidarite	105	94	129	73%	Noms
sont	4	18	38	47%	Auxiliaires être et avoir (mots outils)
sous_officiers	8	4	4	100%	Formes non reconnues
souvenir+	**24**	**85**	**174**	**49%**	**Noms**
strategie	7	6	8	75%	Noms

Figure 62 – Attribution de "souvenir" à la classe "AC"

En prenant comme exemple la déclaration de l'Union nationale des combattants de Saint-Christoly[230] (Gironde), une partie de son vocabulaire – qui comprend les termes « mémoire » et « souvenir » – relève du cluster « Mémoire » dans Calliope alors qu'il est classé dans la classe « Anciens combattants » dans Alceste, dans la mesure où son type de discours est caractéristique de celui des anciens combattants de cette année-là (défense des intérêts matériels et moraux des anciens combattants, maintien des liens de camaraderie et de solidarité, etc.), comme nous le voyons de manière archétypale avec l'exemple ci-dessous :

individu : 191 ** *ID_687 *dpt_Gironde *type_ann_1**
union nationale des combattants de saint christoly, unc saint christoly de blaye, maintenir, dans l' interet superieur du pays, les liens de camaraderie, d' amitie et de solidarite qui existent entre ceux qui ont participe a la defense de la patrie, notamment ceux qui ont vocation a relever de l' organisme officiel en charge des anciens combattants et victimes de guerre; elle peut adherer, en accord avec u n c 33, a un comite d' entente a caractere exclusivement communal; defendre, par tous moyens en son pouvoir, les interets moraux, sociaux et materiels de ses adherents et de leurs ayant droit, ascendants, descendants, conjoints, orphelins; perpetuer, dans la france metropolitaine, dans les departements d' outre mer et dans les territoires d' outre mer, comme chez nos allies, le souvenir des combattants morts pour la france, de servir leur memoire, d' entretenir et de developper des relations fraternelles avec les anciens combattants des nations amies ou alliees.

[230] Le texte de la déclaration, créée en 2010, est le suivant : Titre : « Union nationale des combattants de Saint-Christoly (UNC de Saint-Christoly-de-Blaye) » et Objet : « Maintenir, dans l'intérêt supérieur du pays, les liens de camaraderie, d'amitié et de solidarité qui existent entre ceux qui ont participé à la défense de la Patrie, notamment ceux qui ont vocation à relever de l'organisme officiel en charge des anciens combattants et victimes de guerre ; elle peut adhérer, en accord avec UNC 33, à un Comité d'Entente à caractère exclusivement communal ; défendre, par tous moyens en son pouvoir, les intérêts moraux, sociaux et matériels de ses adhérents et de leurs ayant droit (ascendants, descendants, conjoints, orphelins) ; perpétuer, dans la France métropolitaine, dans les Départements d'Outre-Mer et dans les Territoires d'Outre-Mer, comme chez nos Alliés, le **souvenir** des combattants morts pour la France, de servir leur **mémoire**, d'entretenir et de développer des relations fraternelles avec les anciens combattants des Nations amies ou alliées. » [Nous soulignons]

Or, de même que « commémoration » était associé à la fois, via ses cooccurrents et le type de discours dans lequel il apparaissait, aux deux univers que sont les domaines culturel et militaire, « souvenir » entretient des liens à ces deux domaines, au sein des résultats de Calliope et Alceste.
Nous voyons, au sein du cluster « Mémoire », auquel appartient « souvenir », figurer les termes « solidarité » et « guerre », typiques des déclarations d'anciens combattants ; et « souvenir » est lié (de façon externe), pour plus d'un quart de ses relations lexicales, à « organisme officiel » (qui représente l'Office national des anciens combattants et victimes de guerre, ONAC) :

- ORGANISME OFFICIEL ~ 26,0%
- CEREMONIE ~ 23,4%
- MILITAIRE ~ 12,7%
- VICTOIRE ~ 5,8%
- HONNEUR ~ 5,6%
- CITOYENNETE ~ 5,0%
- COMPAGNON ~ 4,4%
- AUDIOVISUEL ~ 2,6%
- ECHANGE ~ 2,5%
- AINE ~ 2,3%
- EDUCATION ~ 2,2%
- RESISTANCE ~ 1,6%
- REGIMENT ~ 1,4%
- LIBERTE ~ 1,3%
- PUBLICATION ~ 1,0%
- STAGE ~ 1,0%
- CONSERVATION ~ 0,8%
- MUSIQUE ~ 0,5%

Inversement, dans Alceste, si le terme « souvenir » est analysé et classé dans l'univers de discours caractérisant les « anciens combattants », il est néanmoins relié à des termes du domaine culturel. Ainsi, figurent dans son cluster les termes (en bleu) de : « mémoire », « patrimoine », « culture », etc., mais en nombre et en proximité plus faibles que ceux du domaine militaire (en rouge) :

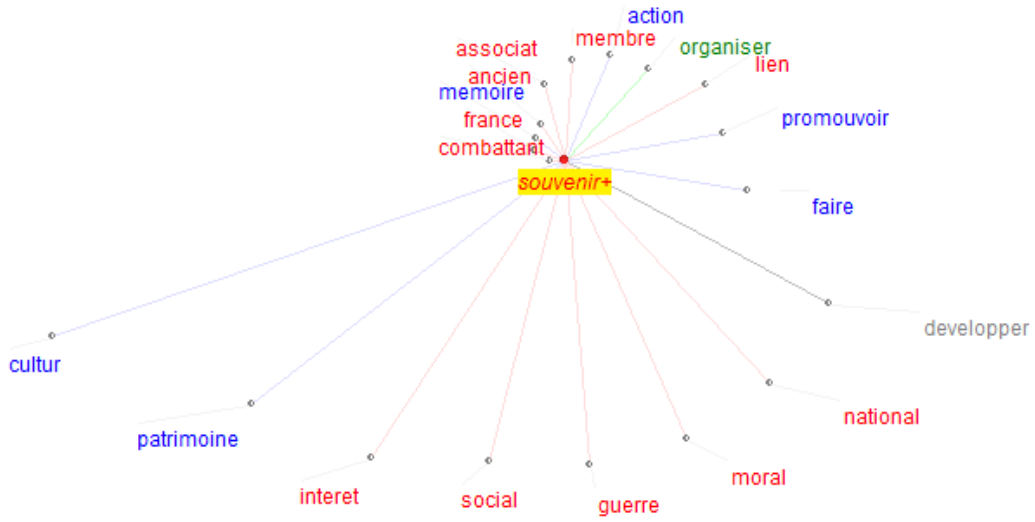


Figure 63 – Réseau lexical du terme « Souvenir+ » dans le corpus

C'est pourquoi, lorsque d'autres occurrences de « souvenir » apparaissent dans des contextes considérés comme culturels par Alceste (sauvegarder, mettre en valeur, photos, films, archives, livres, etc.), ils ne sont pas considérés comme typiques de ce contexte (« souvenir » en gris sur l'image ci-dessous, entouré du bleu qui catégorise la classe « Mémoire », culturelle) :

Unité n° 990 Classe 2 Khi2 = 0 Individu n° 380 *ID_1236 *dpt_LoireA *type_ann_1
omnibus nantes regrouper tous les amateurs de transports urbains de l' agglomeration nantaise: tous les passionnes de transports en commun terrestre: omnibus, tramways, trolleybus, autobus, autocars, vehicules de service, etc;
Unité n° 991 Classe 2 Khi2 = 5 Individu n° 380 *ID_1236 *dpt_LoireA *type_ann_1
reunir, transmettre, sauvegarder, restaurer et mettre en valeur tous les elements de la memoire des transports en commun de l' agglomeration nantaise: photos, films, archives administratives, livres, revues et journaux traitant du sujet, souvenirs oraux retranscrits, petits materiels, pinces, obliterateurs, tickets, machines delivrant des tickets, etc, tenus, maquettes, materiels roulants;

4.3 Complémentarité des deux outils

Pour mettre en valeur la complémentarité entre les deux outils, nous avons choisi certains termes, de Calliope, qui montrait une variation sur la période pour ensuite analyser leurs contextes d'apparition dans Alceste. De cette façon, nous avons pu conjuguer l'intérêt des deux outils : saisir comment évoluait le lexique par le repérage des termes, grâce à Calliope, en lien à leur environnement lexical, avec Alceste.
Ainsi, nous avons vu que des termes en lien aux anciens combattants déclinaient (prisonnier) tandis que d'autres restaient stables (victimes "de guerre") sur la période :

Concordancier de **prisonnier** dans le corpus

association cantonale des anciens combattants, **prisonniers** de guerre et combattants d' algerie, tunisi
association cantonale des combattants **prisonniers** de guerre et combattants d' algerie, tunisi
association des combattants **prisonniers** de guerre et combattants d' algerie, tunisi
association des anciens combattants, **prisonniers** de guerre, sto, rhin et danube, veuves, the
section cantonale de beaumesnil des combattants **prisonniers** de guerre, combattants algerie, tunisie, ma
section cantonale de beaumesnil des combattants **prisonniers** de guerre, combattants algerie, tunisie, ma
e la guerre association des anciens combattants **prisonniers** de guerre et combattants d algerie tunisie
association departementale des combattants **prisonniers** de guerre, combattants d' algerie tunisie m
association depart des anciens combattants et **prisonniers** de guerre du nord combattants d' afn algeri
ale de pecquencourt association des combattants **prisonniers** de guerre combattants d' algerie tunisie ma
association departementale des combatttants, **prisonniers** de guerre, combattants algerie, tunisie, ma
association des anciens combattants, **prisonniers** de guerre, acpg, combattants d' algerie, tu
section locale des anciens combattants **prisonniers** de guerre et combattants d' algerie, tunisi
baisants association des anciens combattants et **prisonniers** de guerre de saint_jean_des_baisants sectio
baisants section locale des anciens combattants **prisonniers** de guerre et combattants d' algerie, tunisi
association des anciens combattants **prisonniers** de guerre, combattants algerie, tunisie, ma
ants d' afn association des anciens combattants **prisonniers** de guerre, combattants algerie, tunisie, ma
te des anciens combattants, victimes de guerre, **prisonniers** de guerre, resistants, combattants algerie,
te des anciens combattants, victimes de guerre, **prisonniers** de guerre, resistants, combattants algerie,
ancien titre association des **prisonniers** de guerre association departementale des an
ociation departementale des anciens combattants **prisonniers** de guerre et combattants d algerie tunisie
ux guerres anciens combattants mutiles reformes **prisonniers** et veuves de guerre association stephanoise
tre association amicale des anciens combattants **prisonniers** de guerre et combattants d algerie tunisie
rtefontaine association des anciens combattants **prisonniers** de guerre et combattants d algerie tunisie
acpg_catm anciens combattants **prisonniers** de guerre_combattants algerie_tunisie_maroc
isie_maroc grouper tous les anciens combattants **prisonniers** de guerre, les combattants d' algerie, tuni

Le terme « prisonnier » apparaît au sein d'énoncés qui désignent souvent les anciens combattants d'Afrique du Nord tandis que celui de « victime » s'intègre dans des énoncés en lien à des catégories plus diversifiées d'associations d'anciens combattants. De plus, ces énoncés ont un vocabulaire plus varié, moins formalisé, stéréotypé que les premiers : il ne s'agit plus seulement de nommer tous les « types » d'anciens combattants (anciens combattants, prisonniers de guerre et combattants d'Afrique du Nord) mais aussi de désigner d'autres victimes des combats (orphelins, victimes civiles et militaires, pupilles, ayants-droits). Il faudrait vérifier, par un retour aux textes, si les associations mentionnées relèvent de fédérations nationales identiques (ce qui indiquerait une évolution lexicale au sein de la fédération) ou différentes (ce qui indiquerait que d'autres fédérations se sont implantées dans le paysage des anciens combattants, avec un lexique différent).

Si nous résumons ce qui vient d'être dit, les deux outils présentent donc différentes vues du même corpus, mais néanmoins concordantes. Nous avons essayé de montrer ici que derrière certaines divergences, nous pouvions retrouver une cohérence entre résultats, car issus du même corpus ! Les résultats obtenus ne sont que des perspectives sur le corpus, dont la multiplicité en dévoile la richesse. Et c'est précisément cette différence entre vues qui constitue l'intérêt de chaque outil. Multiplier les démarches et les outils permet de compenser les points faibles de l'un par les atouts de l'autre, ou plutôt de sortir d'un cadre interprétatif pour en aborder un autre, en vue d'une exploration plus complète de la complexité du corpus.

4.3.1 Calliope : une vision synthétique des thématiques…

L'un des points forts de Calliope réside dans sa capacité à transmettre rapidement un résumé des thématiques traitées dans le corpus. Héritier de la recherche documentaire, il élabore en effet un réseau des termes (descripteurs) cooccurrents dont il donne une représentation synthétique au moyen d'un diagramme stratégique. Celui-ci permet de lire aisément les thèmes du corpus et d'en saisir les éléments capitaux ou, au contraire, secondaires.

En regardant le diagramme stratégique de l'année 1984 ou celui de l'année 2000, par exemple, nous saisissons d'emblée[231] les diverses thématiques qui caractérisent chaque sous-corpus. Le positionnement des clusters dans les quadrants nous indique par ailleurs l'importance (ou non) des thématiques au sein du corpus :

Diagramme stratégique – 1984 **Diagramme stratégique – 2000**

Figure 64 – Diagrammes stratégiques (1984, 2000)

Ainsi, l'année 1984 traite principalement de l'univers des anciens combattants (ayants-droits, liens de solidarité[232]) ; l'année 2000 est représentée par des thèmes plus symboliques (patrie, paix) dont certains émergent (honneur[233]). Les termes « juif » et « politique » font partie des thèmes structurés mais périphériques.
Une analyse de la dynamique temporelle affectant les lexiques peut également être faite à partir de la comparaison de ces deux diagrammes, en étudiant l'évolution du positionnement des clusters au sein des quadrants[234] : ainsi, l'année 1984 voit l'émergence du terme « commémoration » (situé en bas à gauche) confirmé en 2000 (ce terme appartient au cluster « Paix », l'un des plus fédérateurs au cluster). De même, l'accroissement des thématiques culturelles, initiée en 1984 à travers le cluster « Mémoire », se perçoit en 2000 avec le déplacement vers sur la droite du diagramme du cluster « Art et création » (qui contient mémoire).

L'avantage de cette vision synthétique se voit néanmoins nuancé lorsqu'il s'agit d'analyser les multiples liens entre termes. La clarté initiale du diagramme stratégique tend en effet à se brouiller dès que l'on aborde le détail des nombreux liens internes et externes des mots des clusters… au point qu'à la fin, tout semble relié à tout (dans des proportions diverses, il est vrai) ! Le risque existe de recréer soi-même un « discours », qui enchaînerait les différents éléments selon une logique éloignée de celle des documents. Un retour systématique aux sources s'impose donc, comme pour tout logiciel d'analyse de données.
Mais il faut préciser que, dans ce type de démarche, apprendre à discerner l'information pertinente (pour l'analyse) au sein de la multitude de résultats requiert du temps. Il est possible que ce que nous considérons comme une limite de Calliope ne constitue, après tout, que le reflet de notre propre manque de maîtrise !

---

[231] La taille réduite des schémas reproduits ici ne facilite malheureusement pas la lisibilité de ces informations…
[232] Ces deux termes sont en effet positionnés dans le quadrant regroupant les thèmes phares du corpus.
[233] La position de ce terme dans le quadrant sud-ouest traduit l'émergence d'un thème.
[234] « En comparant deux cartes représentant deux flux d'information successifs d'un même domaine de recherche, on prend connaissance de la variation du contenu des agrégats ainsi que de leur migration sur les diagrammes stratégiques successifs, M. de Saint-Léger, *op. cit.*

Par ailleurs, Calliope ne possédant pas de concordancier, la contextualisation des termes passe nécessairement par un retour aux textes. Cette absence peut gêner l'interprétation des résultats dans la mesure où, lorsque les textes sont nombreux, le retour aux textes peut s'avérer fastidieux tandis qu'un concordancier permet très rapidement de visualiser l'ensemble des contextes dans lesquels un terme apparaît. Cela concerne particulièrement les clusters fédérateurs, dont les termes indexent le plus grand nombre de documents.

4.3.2 … doublée d'une analyse chronologique du lexique

L'autre point fort de Calliope – et qui constitue, à nos yeux, son intérêt primordial – est l'opportunité qu'il offre de procéder à une analyse réellement chronologique du corpus. Nous avons vu que l'analyse des clusters au sein des diagrammes stratégiques permettait déjà de révéler les thèmes émergents ou déclinants – outre les thèmes dominants ou mineurs. Cette lecture est complétée[235] par les fonctionnalités de l'outil « Tendances » qui donne à voir, aisément, la dynamique des lexiques dans le temps, au moyen de listes (termes stables, émergents, déclinants ou fluctuants sur une période) et de diverses représentations graphiques (courbes, histogrammes et radars de termes). Certains termes, non immédiatement visibles au sein du diagramme stratégique, sont apparus comme émergents grâce à ces fonctionnalités. Il nous a été en effet possible, non seulement d'analyser finement les glissements qui s'opéraient entre les mots-clés de la question mémorielle, mais également de repérer de nouveaux termes, à faible fréquence mais à poids informationnel croissant, qui en ont permis la représentation.

Nous pensons notamment au thème de l'action sociale, émergeant de façon significative en 2010 (et dont certains termes, comme « reconversion professionnelle » ou « action sociale », apparaissent dans le cluster le plus fédérateur de cette année-là, « Victoire ») :

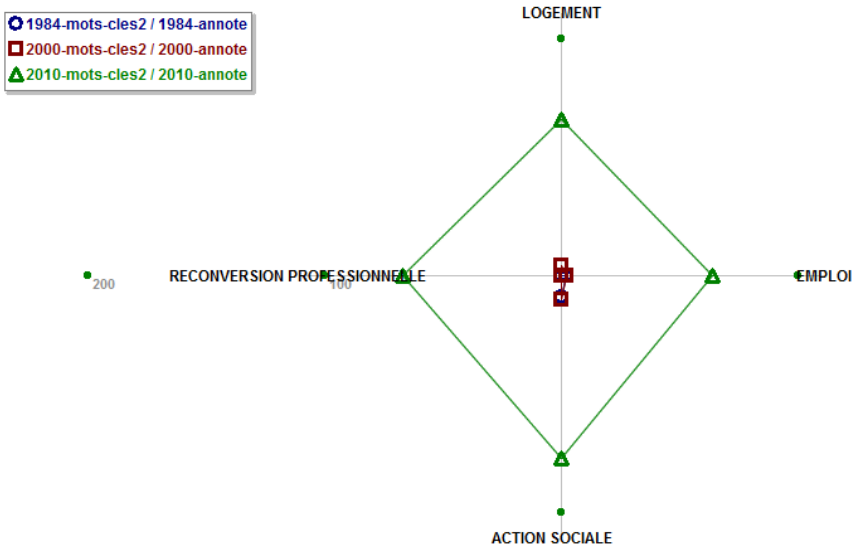


Figure 65 – Termes émergents de la thématique « sociale »

Ou, à certains termes du domaine militaire qui connaissent une régression au cours de la même période :

[235] Cette fonctionnalité résulte en fait de "l'automatisation de la comparaison entre le contenu des quadrants respectifs de deux périodes successives.", *Ibid.*

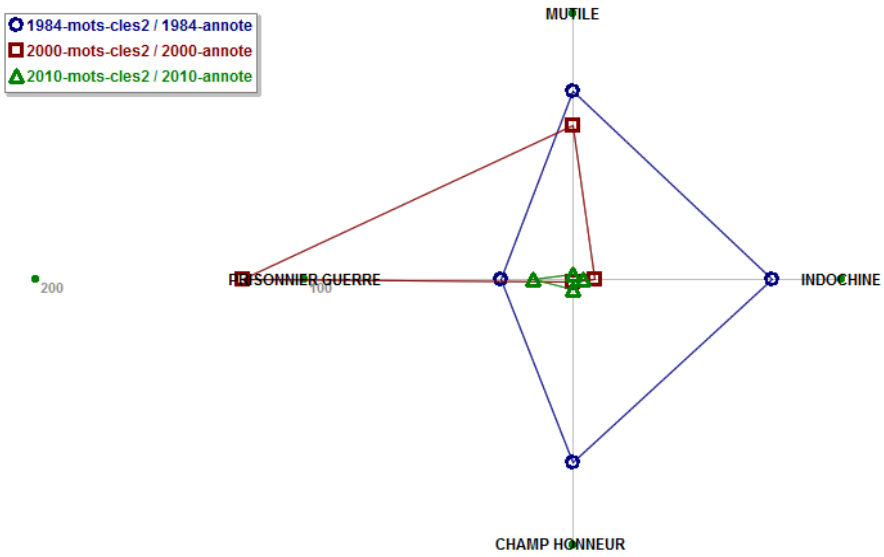


Figure 66 – Termes déclinants de la thématique « militaire »

Ou, enfin, aux thématiques dites « communautaristes », dont voici l'évolution temporelle :

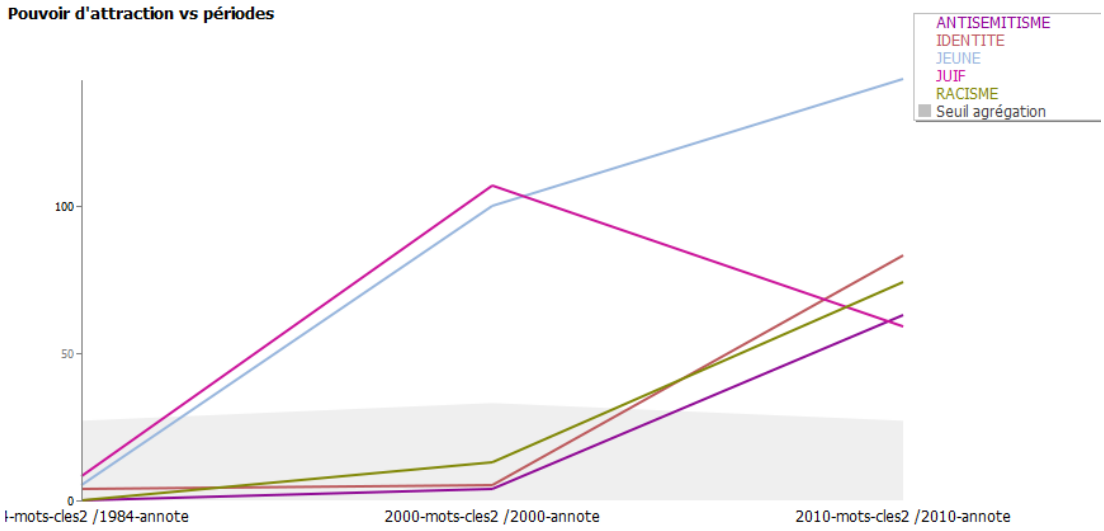


Figure 67 – Evolution du poids informationnel des termes « communautaristes »

C'est précisément parce que Calliope nous a dévoilé, au sein du corpus, l'apparition significative de ces thèmes, que nous avons voulu en vérifier la position (non immédiatement visible) au sein des classifications effectuées par Alceste. Cette capacité de Calliope à permettre une analyse chronologique fine du lexique constitue, pour [39, JENNY], un atout important par rapport à d'autres logiciels. Ainsi, « Comme ce sont les usages des mots dans le corpus, autrement dit leur "profil d'association", qui constituent leur définition – comme dans la tradition des "concordanciers" – on peut repérer certains termes qui changent de signification selon le contexte. Ce résultat d'analyse, très intéressant, a déjà été mentionné à propos du logiciel québécois SATO (Bourque et Duchastel, 1995) et s'oppose aux effets pervers des indexations non contrôlées, comme ceux que peut produire le logiciel Tropes par exemple, en faisant appel à un dictionnaire de référence a priori pour catégoriser les domaines thématiques de tous les mots. »
Il va sans dire que les informations obtenues au moyen de l'analyse chronologique doivent ensuite être complétées par une analyse détaillée des liens lexicaux et du contexte d'apparition des termes (que permet finement Alceste) ; mais ils offrent une première

indication remarquable quant à des éléments potentiellement pertinents pour l'analyse du corpus.

Pour conclure, notons deux petits regrets néanmoins, concernant les fonctionnalités de recherche de Calliope, qui limite le retour aux textes et la vérification des résultats : l'impossibilité d'interroger le corpus par des termes qui n'en seraient pas les descripteurs[236] (ie, les termes qui indexent les textes composant le corpus) ainsi que l'absence de l'opérateur booléen « SAUF[237] », qui permettrait de sélectionner les documents de façon discriminante, et de retrouver des documents contenant tel(s) terme(s), à l'exclusion de tel(s) autre(s).

4.3.3 Alceste : une analyse fine et contrastée des types de discours

En ce qui concerne le logiciel Alceste, son principal intérêt[238] réside dans la finesse d'analyse qu'il permet des différents types de discours qui composent les corpus. Et ce, en raison de deux caractéristiques.

Tout d'abord, la qualité des outils linguistiques « embarqués » (lemmatiseur, traitement morpho-syntaxique du lexique permettant une catégorisation grammaticale fine) font d'Alceste, non pas un logiciel d'analyse du discours au sens strict du terme[239], mais un outil d'analyse de données textuelles à forte orientation linguistique. Les éléments langagiers sont ainsi identifiés et ventilés selon des catégories plus détaillées que dans Calliope. Elles se répartissent en mots pleins (verbes ; noms ; adjectifs ; adverbes ; verbes modaux ; noms propres ; nombres ; lieux et pays ; mois et jours ; couleurs, etc.) et mots-outils (auxiliaires être et avoir ; certains adverbes ; démonstratifs, indéfinis et relatifs ; marqueurs d'intensité ; marqueurs d'une relation temporelle ; marqueurs d'une relation discursive ; marqueurs d'une modalisation ; marqueurs de la personne). Cette finesse dans la catégorisation grammaticale, qui n'existe pas sous Calliope, rend ainsi possible de distinguer les univers de discours par les thématiques traitées (mots pleins) et par la manière de s'y rapporter (mots-outils)[240]. Nous avons ainsi pu mettre en valeur, au sein des diverses classes analysées, les distinctions d'ordre grammatical qui les caractérisaient. Telle classe se distinguait d'une autre par son emploi plus important des verbes (« Patrimoine »), et telle autre par son utilisation singulière des noms de lieux et de pays (« Anciens combattants ») ou de dates (« Commémoration »). Ces éléments d'information donnent de la profondeur à l'analyse, en permettant de préciser le type de discours à l'œuvre.

---

[236] Ceci est peut-être l'indication d'une validation du vocabulaire d'indexation défectueuse.

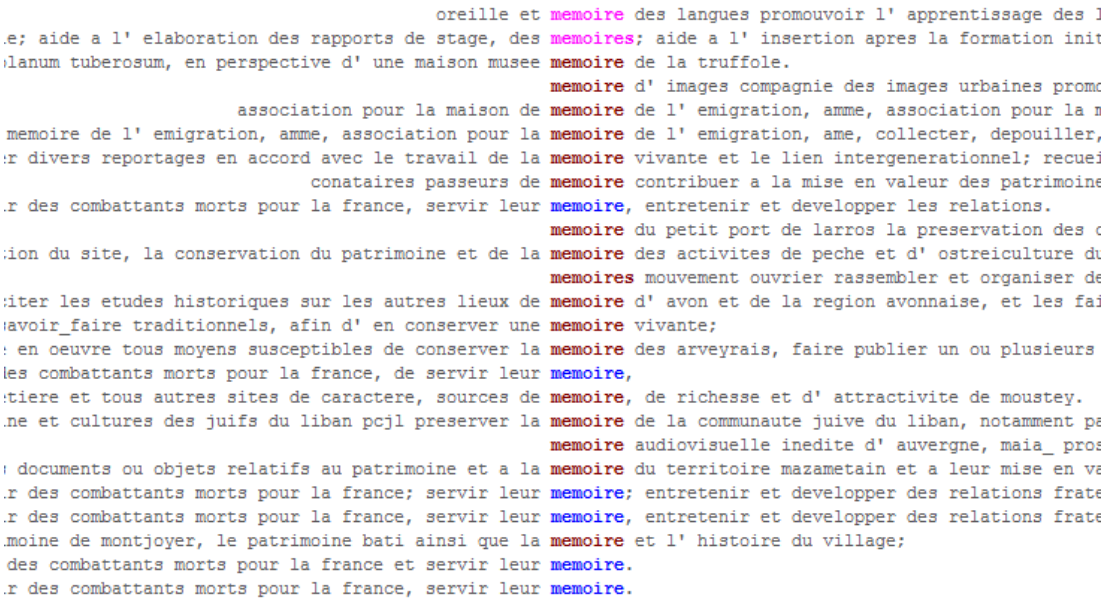
[237] Les opérateurs « ET » et « OU » sont par contre présents. Une plus grande finesse des requêtes est donc possible, relativement à Alceste dont la fonctionnalité « recherche » ne comporte aucun opérateur booléen.

[238] Autre élément intéressant, non développé ici : la production de rapports, détaillé et de synthèse, qui constituent une aide précieuse à l'interprétation des résultats, par une présentation des divers éléments participant à l'analyse.

[239] Le « lissage » lexical opéré par Alceste, dans le processus de lemmatisation (verbes réduits à leur forme infinitive, substantifs à leur forme canonique) peut constituer une limite à une véritable analyse de discours. Précisons qu'il est néanmoins possible de procéder à une analyse du corpus sans lemmatisation. Manque par ailleurs une analyse automatique de la tonalité du discours (ie, le repérage automatique de la polarité d'une opinion par exemple : positive / négative / neutre) ; voir à ce sujet, BOULLIER D. et A. LOHARD, *Opinion mining et Sentiment analysis*, *op. cit.*. Une analyse de la tonalité peut tout de même être réalisée à l'aide des résultats d'Alceste, en étudiant les fragments de texte ou la distribution des différents mots-outils (marqueurs d'intensité, marqueurs d'une relation discursive, etc.).

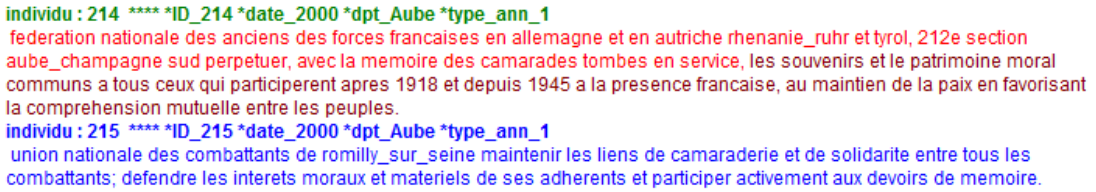
[240] Précisons que si les mots-outils n'interviennent pas dans l'élaboration des classes, le lexique de chaque classe peut néanmoins être analysé au moyen de leurs catégories grammaticales.

Cette analyse linguistique fine du vocabulaire constitue, en quelque sorte, le pendant grammatical de la caractérisation des documents du corpus par leurs fragments les plus typiques. La finalité du traitement opéré par Alceste est effectivement d'établir une typologie des énoncés, une classification des diverses voix qui traversent le corpus, qui s'entrelacent au sein des documents. Comme nous l'avons vu, la révélation de cette structure discursive provient de la méthode de classification descendante hiérarchique qui procède par distinction contrastive, d'étape en étape, entre unités de contexte. Ainsi, les présences caractéristiques d'une classe se transforment, selon un mouvement inverse, en absences significatives d'une autre. Se dévoilent alors, non pas tant des thématiques plus ou moins liées à l'échelle du corpus, que des types de discours en opposition dont les particularités définissent l'appartenance privilégiée à une classe. C'est pourquoi des fonctionnalités telles que concordancier ou segments répétés sont importantes dans Alceste. Elles permettent en effet, en replaçant les termes dans leur contexte d'origine, de saisir les nuances discursives que ces contextes peuvent exprimer, comme nous le voyons ci-dessous avec le terme « mémoire » (issu ici du corpus général de trois années) :



Bien que le terme « mémoire » relève davantage, selon la classification d'Alceste, du type de discours « Patrimoine » (en marron sur l'image), il est possible, grâce au concordancier, de visualiser les autres contextes dans lesquels il apparaît (de manière moins spécifique). Des nuances à l'interprétation peuvent alors être apportées : ainsi, des associations d'anciens combattants (« combattants morts pour la France et servir leur mémoire ») ou d'accompagnement scolaire (« élaboration de mémoires, aide à l'insertion ») emploient le terme « mémoire » dans une finalité non patrimoniale. De cette manière, il est possible de saisir la polysémie des termes qui prennent des nuances différentes en fonction de leur contexte d'apparition.

De même qu'un terme peut figurer dans des contextes relevant de différentes classes, une déclaration d'association peut comprendre à l'intérieur son contenu des énoncés (UCE) relevant de différents types de discours, comme nous le voyons ci-dessous :



Afin de s'assurer de la stabilité des classes définies, Alceste effectue, comme nous l'avons vu en première partie, une double classification en faisant varier la longueur des unités de contexte. Cette répétition différenciée du processus de classification procure ainsi aux résultats une plus grande fiabilité, en limitant les risques de découpage non pertinent.

Les limites d'une telle démarche résident, précisément, dans ce qui en constitue l'intérêt, ie dans la distinction, riche, qu'elle opère.
Tout d'abord, le risque de considérer les classes de discours distinctes comme désignant des documents également disjoints au sein du corpus n'est pas négligeable. Ces types de discours s'appliquent, en fait, à l'intégralité des documents et non à certains d'entre eux (en ce sens, Alceste n'opère pas de typologie de documents) ; de même que les classes d'individus, dans le cadre d'analyse de questionnaires, représentent des sujets épistémiques[241] et non chaque interviewé lui-même.
Ce découpage peut influencer par ailleurs le sens donné aux différents fragments textuels s'ils ne sont pas rapportés à l'ensemble textuel dont ils sont extraits. Ainsi, si l'on étudie l'extrait ci-dessous, nous constatons qu'Alceste a distingué le discours propre aux anciens combattants (en rouge) de celui de la classe « Commémoration » (en vert). Ces deux classes relèvent même de deux branches distinctes de l'arbre – signifiant une plus grande distance lexicale et hétérogénéité entre leurs discours.

individu : 46 **** *ID_46 *date_2000 *dpt_Aisne *type_ann_1
amicale des anciens de la france combattante du departement de l' aisne a l' exclusion de tout but politique ou confessionnel, maintenir les liens d' amitie et d' entraide entre ses membres; satisfaire au devoir de memoire de ce-que fut notre action durant l' occupation de notre patrie, role assure anterieurement par la federation des amicales des reseaux de la france combattante, dissoute au 31 decembre 1999.

Or, l'expression « devoir de mémoire », qui figure dans la partie verte, n'est, de ce fait, pas considérée comme spécifique des anciens combattants – bien que cet extrait provienne d'une association d'anciens combattants. Si la méthode suivie par Alceste peut aider à clarifier des discours parfois mêlés, elle peut, à l'inverse, constituer une source d'erreur dans l'interprétation si l'on ne vérifie pas le texte d'origine. Cette méthode, exigeante, oblige donc à la prudence dans l'interprétation des résultats.

Les analyses d'Alceste nous offrent donc une vision complexe des documents, au sens étymologique du terme, à savoir « ce qui est fait d'éléments différents, imbriqués » que l'analyste aura soin de toujours recontextualiser.

L'autre revers de la médaille, l'autre limite à la classification d'un corpus en énoncés spécifiques, tient justement à ce que seuls les points saillants du corpus sont perceptibles dans l'analyse. Alceste ne retient pas, lors de son traitement, les mots trop partagés par l'ensemble des textes (mots-outils) ni ceux qui sont trop faibles (fréquence inférieure à 4). Si la raison d'une telle manière de procéder se comprend aisément – ce qui est trop ou trop peu présent ne peut prétendre discriminer les discours –, certains éléments pertinents risquent néanmoins de disparaître. Il en aurait été ainsi, pour notre travail, si nous n'avions été alertée par les résultats obtenus précédemment avec Calliope : certains termes, en lien notamment aux thématiques « communautaristes », ne sont pas affectées à une classe par Alceste (comme montré dans l'exemple ci-dessous, avec le terme « esclavage », en gris[242]), ou si c'est le cas, les résultats n'en sont pas immédiatement visibles et demandent à être recherchés.

---

[241] Cette notion, utilisée par Pierre Ratinaud (RATINAUD P., « Outils informatiques appliqués aux sciences de l'éducation »), est issue de l'œuvre de Jean Piaget, qui définit le sujet épistémique comme « l'ensemble des structures d'actions ou de pensée communes à tous les sujets », par opposition au « sujet individuel, qui utilise ces instruments de connaissance ». Pour plus d'informations, voir notamment : http://www.fondationjeanpiaget.ch

[242] Le terme « esclavage » apparaît en gris dans la liste des termes identifiés par Alceste, ce qui signifie qu'il n'est affecté à aucune classe. Sa fréquence totale (7) est supérieure à la fréquence minimale mais il n'apparaît suffisamment en lien à aucun des vocabulaires spécifiques des différentes classes.

Dictionnaire des formes réduites du corpus

Forme	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5	Non classées	Corpus	Marqueur grammatica
entendre	2	0	4	4	0	2	12	Verbes
entente	39	1	0	1	0	8	49	Noms
entraider	53	6	10	6	13	21	109	Verbes
entreprendre	0	0	5	2	2	6	15	Verbes
entreprise	0	0	1	10	7	10	28	Noms
entrer	2	0	0	1	0	3	6	Verbes
entretenir	105	98	27	19	13	31	293	Verbes
entretien	1	0	12	9	5	7	34	Noms
environnement	0	0	1	24	2	10	37	Noms
environs	33	5	4	13	2	15	72	Noms
epanouir	0	0	2	0	5	0	7	Verbes
epoque	0	0	1	6	5	7	19	Noms
equilibre	0	0	0	1	3	3	7	Noms
equipement	0	0	0	8	1	5	14	Noms
eriger	2	0	8	1	0	0	11	Verbes
esclavages	**0**	**0**	**1**	**2**	**0**	**4**	**7**	**Formes reconnues mais**
espace	0	0	2	17	9	15	43	Noms

Figure 68 – Distribution de la fréquence de la forme « esclavage »

4.3.4 Une analyse chronologique « naturellement » spécifique

Enfin, dernière précision, l'analyse chronologique qu'offre Alceste a également les défauts de son avantage. Il est en effet possible d'introduire, dans le corpus, une variable temporelle illustrative (« date » par exemple), qui permet d'en croiser les résultats : Alceste identifiera alors un type de discours particulier comme caractérisant davantage une année – en raison de sa surreprésentation cette année-là. Ces résultats, importants, ne correspondent néanmoins pas à une analyse chronologique qui serait réellement dynamique : la variation des termes au fil de la période étudiée n'est pas mise en avant, les glissements lexicaux doivent être repérés un à un, contrairement à ce que permet Calliope.

Si nous devions indiquer une petite limite de cet outil, nous dirions que le retour aux textes, de façon « orientée », n'est pas toujours aisé. Par « orientée », nous entendons le retour aux textes initiaux, complets, à partir de certains résultats, présentés dans les graphiques. Dans Alceste, l'interrogation des textes ne se fait qu'à partir de l'écran « Unités textuelles du corpus », les autres visualisations de résultats renvoyant aux seuls segments de texte ; de plus, les requêtes sont sommaires, limitées à un seul terme (absence d'opérateurs booléens). Cette particularité gêne parfois la vérification d'un résultat.

4.3.5 Analyse des exclusions opérées par les deux outils

Pour conclure notre comparaison des deux outils, nous avons étudié, outre les résultats produits, ce que ces résultats excluaient. Autrement dit, analyser ce qui ne relève pas de la sphère des résultats devrait permettre, selon nous, de mieux comprendre le fonctionnement des logiciels et d'évaluer la pertinence des résultats mis en avant. De manière générale, certaines exclusions ne semblent pas sujettes à caution : ainsi, Calliope a classé dans son fichier « bruyant[243] » des déclarations dont la taille est trop petite pour être traitée[244] ou

[243] Le fichier bruyant rassemble les documents dont les termes n'apparaissent dans aucun cluster : le vocabulaire de ces documents ne participe donc pas à la construction des clusters. De même, un terme bruyant est absent de tout quadrant du diagramme stratégique.

[244] Par exemple, la déclaration de l'association « UNC Saint-Pierre-de-Plesguen, Association d'anciens combattants », créée en 2000, ne pas contient pas suffisamment de termes pour pouvoir être traitée par Calliope. A noter que la majorité des déclarations exclues pour cette raison concerne des associations d'anciens combattants, qui ne spécifient très souvent que leur simple localisation géographique, adjointe à la mention de la fédération nationale d'anciens combattants à laquelle elles sont affiliées.

dont le contenu est hors sujet[245]. Dans cette dernière catégorie, nous pouvons citer les déclarations suivantes : « Retraite sportive bagneraise. Proposer différentes activités sportives à partir de cinquante ans : randonnées pédestres, vélo, atelier de mémoire » ou « Je me souviens... Soutenir les actions qui visent à promouvoir et accroître l'autonomie des personnes âgées ; prévenir les effets du vieillissement en diffusant les techniques de stimulation des fonctions mnésiques ; assurer dans ces domaines la formation et l'information des différents intervenants auprès des personnes âgées ». Ces deux déclarations se voient écartées du traitement effectué par Calliope en raison de leur emploi du terme « mémoire » dans le sens d'une faculté cognitive individuelle, qui contraste avec le sens dominant de mémoire au sein du corpus, entendue comme « mémoire collective ». Rappelons qu'Alceste a, quant à lui, conservé ces occurrences de mémoire pour les classer dans la catégorie « Echanges », qui insiste sur le développement personnel et les relations interpersonnelles, sur un lien au présent et non au passé.

En analysant les termes considérés comme bruyants par Calliope, nous remarquons qu'ils sont de moins en moins nombreux dans le temps, malgré un nombre de déclarations qui, lui, s'accroît : l'année 1984 en compte 19, l'année 2000, 14 et l'année 2010, 3. Ce qui signifie que le vocabulaire des déclarations est de plus en plus homogène (de moins en moins de termes ne sont pas traités, ie les termes voient leurs liens augmenter au fil du temps). En 1984, les termes bruyants comprennent notamment des mots en lien aux anciens combattants : conscrit, organisme officiel (ONAC), Belfort (« union nationale des anciens combattants d'Indochine, des TOE et d'Afrique du Nord, section de Belfort » ; « amicale des marins et marins anciens combattants de Belfort » ; « fédération nationale des anciens des forces françaises en Allemagne et en Autriche, section 201 du territoire de Belfort »). Ces termes ne sont pas suffisamment présent ni reliés suffisamment à d'autres termes pour participer à la construction de clusters.

Les autres termes évacués de l'analyse appartiennent notamment au domaine patrimonial : chapelle, port, artisanat, terroir, restauration – ce qui confirme l'aspect émergent en 1984 de cette thématique, qui se développera dans les fichiers des années suivantes.

A noter le terme « discrimination » qui n'apparaît qu'une fois en 1984 (d'où son exclusion de l'analyse) mais au sein d'une déclaration d'anciens combattants (« bureau d'aide sociale, office d'anciens combattants, le secours qui leur est nécessaire pour assumer la difficulté du moment, elle prend en charge, dans la mesure de ses moyens, les détresses de toute nature qui lui sont signalées et cela sans à priori ni discrimination »).

En 2000, aucun terme associé aux anciens combattants n'est considéré comme bruyant : leur fréquence et liens aux autres éléments du lexique se sont donc renforcés. Par contre, des termes liés au patrimoine et à l'histoire continuent à ne pas être suffisamment présents et typiques : terroir, breton, Bretagne (« souvenir bourbonien en Bretagne, approfondir la connaissance historique des Capétiens »). Enfin, des termes qui seront davantage présents en 2010 commencent à apparaître (mais pas suffisamment pour être traités) : autonomie, cadre de vie, lecture, compréhension, étudiant.

En 2010, le terme « camelot » est considéré comme bruyant, ainsi que « rue » (employé pour désigner le lieu de la domiciliation de l'association) et « métier ». Ce dernier apparaît en lien avec différents types de discours, selon Alceste :

[245] Par contenu hors sujet, il faut entendre un contenu dont les termes se révèlent trop spécifiques, qui ne cooccurrent pas régulièrement avec le reste du lexique traité.

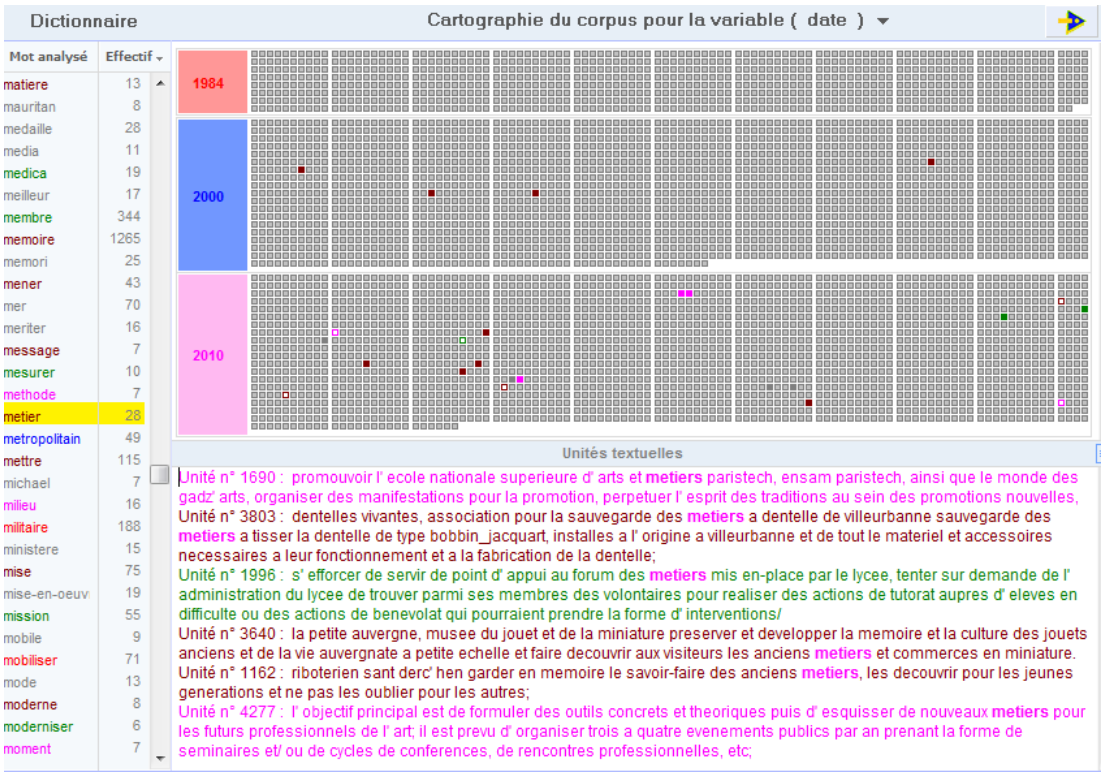


Figure 69 – Distribution du terme « métier » par année

En ce qui concerne Alceste, celui-ci a supprimé de l'analyse les termes partagés par trop ou trop peu d'unités textuelles (les prénoms notamment).
En raison de leur présence trop importante, les mots-outils (« et », « pour », « qui », « entre », « dans », etc.) sont nécessairement supprimés de l'analyse par les deux logiciels. Sinon, ce sont ces mots qui présideraient à l'élaboration des classes et clusters, en raison de leur poids très important dans le corpus. Il est néanmoins possible, dans Alceste, d'analyser la présence de ces termes au sein des divers types de discours, en étudiant leur distribution dans les classes. Dans le cadre de la poursuite de notre étude, nous étudierons donc les marqueurs de la temporalité que sont « hier », « aujourd'hui », « demain », « depuis », « durant », qui peuvent peut-être apporter d'autres informations pertinentes à la thématique mémorielle.
Par contre, d'autres exclusions opérées par ces outils s'avèrent plus surprenantes. Calliope a, par exemple, écarté les déclarations suivantes : « Association franco-britannique de la Plaine-au-Bois. Sauvegarder le lieu de mémoire constitué par la pâture du massacre de la Plaine-au-Bois à Esquelbecq ; rassembler les fonds pour l'acheter ; faire connaître aux visiteurs et touristes l'histoire de ce lieu » et « Office municipal de la mémoire. Collecter, sauvegarder et communiquer les traces ou éléments du passé de Liévin et environs, animation et promotion de journées symboliques en matière d'histoire locale », alors que ni leur taille ni leur vocabulaire n'expliquent ce rejet…
Alceste a, par ailleurs, sorti de l'analyse le terme… « combattant », dont la fréquence s'élève à 3 183 (combattant constitue le 2e terme le plus fréquent du corpus, juste après « et »), bien loin devant les termes « ancien » (1 419) et « mort » (486), auxquels il est souvent accolé. S'il apparaît parmi les premiers segments de la liste des segments répétés (« Union nationale des combattants », « ancien combattant », etc.), le terme lui-même ne participe pas à l'élaboration des classes…

4.4 Multiplier les outils au service d'un projet exploratoire

Nous venons de voir que chaque outil possède un fonctionnement propre qui génèrent des résultats insistant sur des aspects différents mais également intéressants du corpus : Calliope, en offrant une vue synthétique du réseau lexical thématique et en procédant à l'analyse dynamique du vocabulaire ; Alceste, par la richesse et la complexité de son analyse discursive. Le recours aux deux logiciels permet donc de mettre en valeur certains points du corpus et de compléter une analyse par l'autre. Calliope a notamment permis de repérer des tendances temporelles (ce que permet plus difficilement, et différemment, Alceste) et Alceste, de distinguer les différentes voix qui traversent le corpus (qui ne sont pas distinguées au sein de Calliope) et la polysémie des termes.

Mais les véritables enseignements tirés de ces résultats, et plus généralement, de la démarche dont ils sont issus, se situent à un autre niveau : révéler des pistes qui permettent de nourrir le projet de recherche, sur lesquelles asseoir les décisions quant à la poursuite du projet.
Cette étude, rappelons-le, n'avait d'autre but que de procéder à une première exploration de certaines données du *Journal officiel* en vue d'une analyse plus systématique. Exploratoire, donc, elle entendait, non pas fournir des résultats visant à conforter (ou réfuter) les hypothèses de recherche initiales mais à mettre au point une méthode de sondage des données de manière à pouvoir opérer des choix qui en préciseraient à la fois l'exploitation future et le questionnement. Le recours aux outils d'analyse textuelle a ainsi permis de prendre les décisions suivantes :

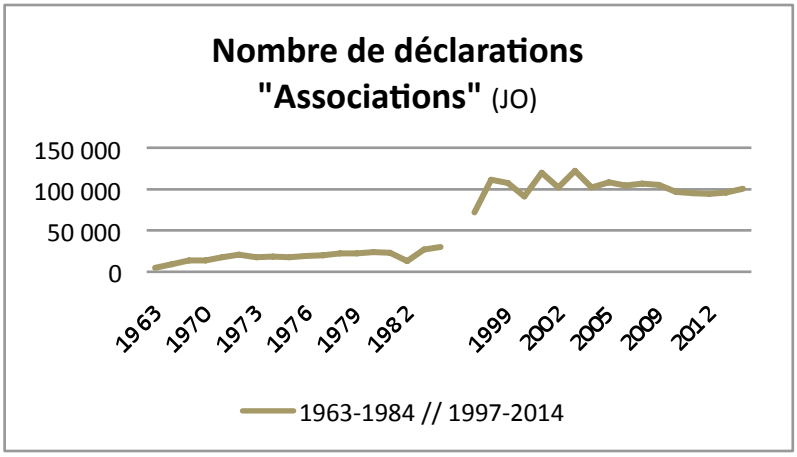
4.4.1 Analyse future des déclarations au JO : conservation du seul mot-clé « mémoire »

La réduction du corpus que nous avons opérée répondait, outre aux contraintes temporelles, à cet impératif exploratoire : nécessairement non représentatif, il s'est limité à traiter trois années (1984, 2000 et 2010) de déclarations, sans distinction de contenu (entre titre et objet des déclarations) ni de mot-clé. Mais l'analyse des résultats a permis de d'orienter le projet, de fermer des portes pour en ouvrir d'autres. Les résultats ont ainsi montré l'inutilité d'analyser les déclarations contenant les quatre mots-clés mémoriels : nous retrouvons en effet en résultats (et cela, de manière particulièrement flagrante avec les classifications d'Alceste) les distributions lexicales propres à chaque mot-clé, qui correspondent de fait à la distinction initialement introduite. Afin d'échapper à cette circularité, rassurante mais non constructive, il a été décidé de ne retenir au sein du futur corpus que les déclarations contenant le mot-clé « mémoire ». Le filtrage au moyen des quatre mots-clés représentait en effet le double inconvénient de ne pas être exhaustif (tous les termes potentiellement représentatifs de la question mémorielle ne sont pas retenus) tout en demeurant trop large. La finalité du projet s'en est trouvée affermie : au lieu d'étudier les lexiques liés aux anciens combattants, au souvenir ou à la commémoration et de voir leur éventuelle jonction avec celui de la mémoire, il s'agira de se focaliser sur ce dernier pour analyser la manière dont il se mêle, se tisse à d'autres thèmes. Le terme « mémoire » doit donc constituer le pivot de l'analyse, autour duquel gravitent, plus ou moins, les autres termes. C'est précisément ce lien de distance / proximité lexicale à la mémoire qu'il s'agira de scruter : dans le temps, dans l'espace, selon le type d'annonce et de contenu.
Les résultats ont également permis de prendre une autre décision : une fois le corpus « mémoire » constitué et analysé, nous procéderons à une seconde analyse mettant de côté le mot-clé lui-même. De cette manière, nous pourrons étudier plus finement la distribution lexicale des thèmes connexes sans la présence de ce terme, au poids évidemment prépondérant puisqu'il aura présidé au filtrage des déclarations.

4.4.2 Elargissement de la couverture temporelle du corpus par le regroupement de déclarations

Autre apport de l'étude : nous avons vu que pour répondre aux postulats statistiques des outils d'analyse de données textuelles, il fallait se doter d'un corpus à la taille suffisamment

importante. Au-dessous d'un certain seuil, le traitement des données n'a en effet statistiquement aucun sens. C'est pourquoi il a été décidé de procéder à des regroupements de fichiers au sein de « paquets » d'années de déclarations, afin que les périodes anciennes étudiées (dont le nombre de déclarations annuelles est beaucoup plus faible que celui des années récentes) puissent, en dépassant ce seuil, être traitées. Concrètement, il s'agira de déterminer le nombre minimal de déclarations contenant le mot-clé « mémoire » pouvant être traité par les outils statistiques. Etant donné l'inflation associative constatée au fil du temps (dont nous reproduisons l'évolution ci-dessous), le seuil à atteindre pourra alors nécessiter de regrouper les déclarations de plusieurs années au sein de paquets.



Une fois ce seuil déterminé[246], le regroupement en x années sera reproduit sur les périodes suivantes, espacées dans le temps d'un certain nombre d'années[247].
Cette manière de procéder rendra possible l'élargissement du périmètre temporel de l'étude, en intégrant des années qui en étaient auparavant exclues, en raison du faible nombre de leurs déclarations.
Outre l'extension temporelle du corpus, cette méthode de « ponctions » groupées permettra une distinction du contenu ou du type d'annonce, au sein des déclarations : une analyse du lexique selon le titre ou l'objet de la déclaration ou selon la création ou la modification est alors envisageable.

4.4.3 Couverture géographique : sélection de départements

Dans la mesure où l'accès aux fichiers du JO archivés à la BnF, pour la période 1985-1993 ne permet pas une exploitation aisée des données (fichiers conservés sous forme de microfiches), il a été décidé de sélectionner certains départements présentant un intérêt particulier pour la thématique de recherche. Il s'agit des départements suivants : Bouches du Rhône (13) ; Cantal (15) ; Isère (38) ; Nantes (44) ; Rhône (69) ; Paris (75) ; Hauts de Seine (92) ; Seine-Saint-Denis (93) ; Martinique (972). Cette sélection permet de rassembler des caractéristiques variées, constituant un échantillon dont la diversité devrait permettre d'analyser les multiples aspects du phénomène mémoriel (département fortement urbanisé ou à dominante rurale, variété des catégories socio-professionnelles des habitants, lien à l'histoire de la résistance, de l'esclavage, à l'immigration, au patrimoine et à la culture, etc.).

[246] Par exemple, afin d'obtenir les 150 déclarations minimales nécessaires au traitement statistique, il faut, à partir de 1963, regrouper les déclarations des 3 années suivantes.
[247] Si nous reprenons l'exemple précédent, les paquets seront constitués de 4 années de déclarations, ponction ensuite répétée tous les 5 ans. Nous obtenons de cette façon une méthode d'échantillonnage, reproductible sur toute la période considérée.

4.4.4 La mention de dates dans les déclarations : un passage obligé de la mémoire ?

Enfin, Alceste a mis en lumière la mention caractéristique de dates au sein de certaines déclarations (notamment, la classe « commémoration »). Une attention sera donc portée à ce fait dans les études futures afin de voir s'il correspond, d'une part, à l'inflation des journées nationales commémoratives relevée par Sarah Gensburger et, de l'autre, à l'instauration d'un nouveau canon de la représentation mémorielle. Le phénomène mémoriel s'exprime-t-il (davantage ?) par la référence à un événement historiquement daté ? Il s'agira donc de vérifier si un tel phénomène existe, d'en mesurer l'ampleur, puis d'en analyser le lexique (quels sont les événements historiques auxquels les dates se réfèrent et de quelle manière les déclarations s'y réfèrent-elles ?), l'évolution temporelle et la répartition géographique.

Conclusion

A l'issue de ce travail, nous aimerions, en guise de conclusion, revenir sur les principaux apports et limites de la démarche suivie dans ce projet, à partir desquels nous proposerons quelques pistes de réflexion. Mais auparavant, nous souhaitons insister sur les enjeux que représente l'analyse de données textuelles, notamment pour les sciences sociales.

Depuis l'avènement des premiers ordinateurs, dans les années 50, des méthodes statistiques et linguistiques ont été appliquées à des matériaux textuels. Que ce soit pour en proposer une traduction automatisée ou pour en extraire des spécificités lexicales, ces méthodes n'ont cessé d'être utilisées, dans les domaines académiques comme dans la sphère privée, pour tenter d'appréhender, de façon automatisée les éléments significatifs – pour ne pas dire signifiants – des documents textuels.
Avec l'accroissement exponentiel des capacités computationnelles des ordinateurs et la diffusion de plus en plus large de leurs usages, la démultiplication de sources de données textuelles (dont les nouvelles sources que sont les réseaux sociaux), le développement d'outils d'analyse et de représentation des données (graphes de réseaux), le partage de données à travers le web sémantique, la volonté d'ouverture des données publiques dans le cadre d'une démocratisation de l'accès à l'information, ce sont autant de transformations technologiques et de modifications sociales qui bouleversent profondément les savoir-faire et les méthodes de nombreux métiers.
La sphère académique n'est pas épargnée. Là où auparavant elle avait affaire à des corpus limités, traités et analysés manuellement, elle se voit désormais confrontée à des masses d'informations qu'aucune lecture intensive ne peut appréhender. Le recours à des techniques informatisées s'avère donc indispensable.
L'analyse de données textuelles constitue justement l'un de ces moyens d'accès « automatisés » aux textes : par ses opérations de réduction des données et de représentation schématique de gros corpus textuels, elle représente une réelle opportunité pour les sciences humaines et sociales – ces sciences dont l'une des matières essentielles est composée de matériaux textuels. Il n'est que de penser aux corpus littéraires – qui, précisément constitueront le premier domaine d'application des outils lexicométriques – ou encore aux bibliothèques numériques de manuscrits grecs ou latins pour s'en convaincre.
A cette capacité de traiter de volumineux corpus s'ajoutent des méthodes de lecture distanciée des matériaux qui permettent de limiter les a priori interprétatifs et les catégorisations préalables. Comme le propose [17, MORETTI], dans *Graphes, cartes et arbres* notamment, à propos de l'histoire de la littérature, il s'agit de « lire autrement les livres, de lire extensivement ou plutôt de voir "de loin" ». Ce nouveau type de lecture gouverné par les données (« data-driven ») permet d'accéder aux différentes formes constitutives des matériaux textuels comme autant d'indices fournis à l'interprétation de l'expert. « La distance n'est pas un obstacle mais une forme spécifique de connaissance : un nombre plus réduit d'éléments, d'où un sens plus aigu de leur interconnexion globale. Organisations, relations, structures. [17, MORETTI] » Les principes statistiques sur lesquels sont fondés les outils lexicométriques opèrent ainsi une mise à plat des données, au sein de tableaux lexicaux, puis évaluent leurs relations (dépendance / indépendance) pour en montrer la structure, la forme au moyen de représentations graphiques. Cette capacité à traiter le tout pour en extraire la structure schématique ou les points saillants constitue, à nos yeux, l'un des principaux atouts de l'analyse de données textuelles.

Et pourtant, selon [13, DACOS ; 31, DEMAZIERE et al.], les sciences sociales manifestent une certaine réticence envers cette démarche, considérée comme éloignant le chercheur des données, à la fois par la lecture distanciée qu'elle en propose mais également par l'accès à d'autres sources qu'elle permet. La proximité entre le chercheur et ses données est en effet une caractéristique des sciences sociales. Non nécessairement contraint comme l'historien ou le philologue à une distance vis-à-vis de ses matériaux d'étude, le sociologue construit le plus souvent ses propres données, à partir de son terrain d'étude. Recueillies sous forme de notes de terrain, de réponses à des entretiens ou à des questions ouvertes, ces données lui fournissent un aperçu du social qu'il a circonscrit et sélectionné, voire même motivé. D'où les

questions épistémologiques et méthodologiques propres à ces disciplines (sociologie, ethnologie, anthropologie, etc.) d'observation participante, d'observation directe, d'étude de terrain, en rupture avec les anciennes pratiques ethnographiques de lecture de récits de voyageurs, de rapports de fonctionnaires ou de journaux de missionnaires. L'éloignement du terrain que représente l'accès à des données élaborées par d'autres, provenant de sources indirectes représente une perte de maîtrise sur l'objet d'étude. « Que devient alors le rapport aux enquêtés si celui-ci ne repose plus sur une proximité – et un contact direct – avec la population mais sur une mise à disposition d'informations dites de "seconde main" ? Par quels moyens la relation d'enquête est-elle maintenue ou simplement annihilée ? Comment aller chercher l'information disponible sur les bases de données et y arrimer des hypothèses de travail pertinentes ? » se demandent [31, DEMAZIERE et al.].
Pourtant, avec l'analyse secondaire des données se profilent des pistes nouvelles d'exploration de la matière sociale. Réexploiter des données construites dans une autre perspective, en vue d'autres problématiques représente un nouveau défi pour les sciences sociales, une démarche novatrice rendant possible de nouvelles interprétations d'un même matériau.

La mise à disposition de données de plus en plus nombreuses sous format numérique ouvre donc des perspectives de recherche intéressantes. Il est dès lors possible d'accéder à de nouveaux matériaux, de nouvelles sources auxquels les outils d'analyse de données prêtent leurs capacités de calcul et de synthèse graphique. En ce sens, ils semblent profondément adaptés aux nouveaux besoins de la recherche scientifique. Et ce n'est pas un hasard si des outils conçus dans les années 80 et 90 (Alceste, Calliope via Leximappe, Hyperbase, Lexico, etc.) connaissent de nos jours un regain d'attention de la part de la communauté académique. Le besoin se fait en effet sentir de pouvoir mieux « cerner » des données dont le nombre et l'extériorité rendent la connaissance « intime » difficile, voire impossible.
Le projet auquel nous collaborons repose précisément sur cette démarche : appliquer des logiciels lexicométriques à des données non élaborées par le chercheur mais collectées en vue d'autres finalités que la recherche scientifique : les déclarations d'associations publiées au *Journal officiel* et diffusées par la DILA. Pour rappel, ce sont plus de 2 500 déclarations qui ont ainsi pu être traitées et analysées dans le cadre de notre stage.
L'accès à une source extérieure conjugué à la manipulation statistique des données constituent une démarche encore originale en sciences sociales de nos jours en France. Cette démarche a pourtant permis de « révéler » certains aspects du phénomène mémoriel, qui peuvent être complétés par une démarche plus classique d'observation sociologique ou d'entretiens avec des membres d'associations, par exemple. Car, s'il faut prendre au sérieux une telle méthode, ce n'est pas pour la promouvoir à l'exclusion d'autres méthodes. Elle permet un accès supplémentaire à la matière sociale, sans prétendre la couvrir intégralement.
De plus, l'application de deux outils lexicométriques au même corpus a permis de proposer des lectures différenciées et complémentaires des matériaux textuels. En effet, bien que leurs résultats soient globalement concordants, chaque outil utilisé, Calliope et Alceste, a mis en avant des aspects particuliers du corpus : Calliope, au moyen d'une analyse véritablement chronologique des documents, a permis de saisir sur la période étudiée des thématiques de la question mémorielle en émergence (notamment les expressions « devoir de mémoire » et « lieu de mémoire ») ou, au contraire, en régression (« mutilé », « prisonnier de guerre » ou « champ d'honneur », termes liés aux anciens combattants) ; Alceste, de son côté, a pu mettre en valeur des types de discours caractérisant des aspects différents du phénomène mémoriel (et notamment, une distinction entre des types de discours propres aux anciens combattants, qui mobilisent davantage le lien au souvenir de camarades morts pour la Patrie, et des types de discours qui caractérisent d'autres acteurs de conflits passés, résistants ou victimes civiles de guerres par exemple, qui privilégient une dimension symbolique et un lien au passé qui se traduit par l'organisation de cérémonies commémoratives). Ces premiers résultats constituent autant de pistes interprétatives dans l'étude de la question mémorielle.
La spécificité des résultats obtenus avec les deux logiciels résulte, bien sûr, de celle qui sépare leurs principes de fonctionnement. Et, c'est précisément cet aspect qui constitue le

point fondamental de notre projet : avoir montré que l'utilisation conjointe d'outils différents constituait un intérêt pour les projets de recherche en offrant différents éclairages d'un même corpus de documents. La multiplicité des points de vue posés sur un même corpus permet ainsi de multiplier les questionnements, de susciter d'autres pistes à explorer. Loin d'être opposées, ces méthodes se sont révélées au contraire profondément complémentaires.

Malgré l'atout certain que peut représenter le recours à l'analyse de données textuelles, en sciences sociales notamment, nous avons vu que des risques existaient néanmoins – risques qu'il faut connaître pour pouvoir mieux les limiter.
Le premier risque serait de considérer les résultats comme spontanément signifiants, voire « objectifs ». L'attraction du chiffre et de l'image, pour séduisante qu'elle paraisse, nécessite, pour la conjurer, de mettre à distance les résultats et de les rapporter à leurs modes d'élaboration. Fondés sur des postulats statistiques et linguistiques, qui sont autant de partis pris sur le matériau textuel, les outils d'analyse de données textuelles ne sont pas neutres. La spécificité des résultats obtenus avec Calliope et Alceste en est la preuve : clusters de termes d'un côté, types de discours de l'autre – résultats différents, pourtant issus du même corpus de données.
Cela ne signifie pas qu'il faille rejeter ces outils au motif qu'ils ne seraient pas « objectifs » (quelle méthode pourrait l'être d'ailleurs ?) mais simplement que leurs principes de fonctionnement doivent être connus et maîtrisés pour tirer profit de leur usage. Nous espérons avoir montré que les opérations appliquées au corpus par ces logiciels déterminaient les résultats, en ce qu'elles définissent ce qui doit être traité, la manière de le traiter et celle de le restituer. Le paradigme du « mot plein » comme représentant canonique de ce qui est dit constitue une hypothèse forte qui ne doit pas être oubliée lors de l'interprétation. De même, le processus de lemmatisation et de catégorisation grammaticale opère des choix sur le matériau textuel qui sont loin d'être anodins et qui influencent les résultats. Nous avons souligné, à ce propos, l'adéquation que représentait notre corpus aux techniques d'analyse de données textuelles : documents formatés, essentiellement descriptifs, les déclarations au JO font état d'un style discursif particulièrement adapté aux deux logiciels lexicométriques retenus, Calliope et Alceste. Surreprésentation des verbes à l'infinitif et des substantifs, faible présence de marqueurs de modalité, ces textes échappent ainsi en partie aux risques inhérents aux outils d'analyse de données textuelles. Une prudence accrue aurait dû être adoptée si d'autres types de documents avaient été soumis à leur traitement – nous pensons notamment à des recueils d'entretiens. Enfin, la signification accordée à la fréquence – même si elle est relative – des termes, à leur distribution et répétition au sein du corpus est à prendre en considération. Selon [26,BASSAC et al.], Alceste constitue d'ailleurs des classes via la répétition de termes « moins dans la mise en discours que dans le repérage de leur simple présence / absence » au sein de tableaux lexicaux. Cette abstraction opérée sur les données définit bien les méthodes statistiques.
Mais rappelons que les statistiques ont une longue histoire et qu'à l'instar de toute élaboration conceptuelle, les formulations statistiques sont le résultat de choix et d'hypothèses. Comme le souligne A. Desrosières, « Les difficultés rencontrées au cours de l'histoire [par les sociologues sont les mêmes que celles rencontrées] par les inventeurs des langages statistiques qui, précisément, permettent de constituer les faits sociaux en choses. Aujourd'hui ces langages s'appuient sur des concepts synthétiques bien formalisés : moyenne, écart-type, probabilité, classe d'équivalence, corrélation, régression, échantillon, etc. L'étudiant, le chercheur ou l'utilisateur de données statistiques reçoivent des concepts compacts, encapsulés dans des formulations concises et économiques, alors que ces outils sont le produit d'une gestation historique traversée d'hésitations, de retraductions, de conflits d'interprétation. Pour les maîtriser, l'apprenti doit se poser et résoudre en peu de temps des questions discutées pendant des décennies ou des siècles. Rouvrir ces débats ne relève pas d'une curiosité érudite, ajoutée comme un supplément d'âme à l'acquisition de techniques formalisées, mais fournit un cheminement et une aide au processus de compréhension et d'apprentissage. […] L'histoire fait comprendre comment les faits sociaux sont devenus des choses, et, par là, comment ils le deviennent pour chaque utilisateur des techniques statistiques. [73, DESROSIERES] »

Comme toute démarche scientifique, les statistiques, notamment textuelles, demandent donc recul et réflexivité. Et ce n'est qu'à ce prix que leurs résultats gagneront leur place dans la construction de la validité scientifique en sciences sociales. Utiliser les méthodes de l'analyse textuelles en sciences sociales relève donc de la même posture, scientifique, à ceci près que l'habillage d'objectivité dont l'habitude a revêtu les premières obligent à un surcroît de distance à leur égard.

Pour limiter les risques propres à l'utilisation d'outils d'analyse de données textuelles, il faut non seulement en comprendre les principes mais aussi intégrer cette démarche dans une démarche plus globale, qui en détermine la finalité et le sens. Seul le questionnement scientifique à l'origine du projet peut et doit guider l'analyse des données. Il est en effet très tentant, lors du processus de traitement ou de la production de résultats, de se laisser dépasser par l'aspect purement technique des outils ou séduire par la force graphique des résultats – qui plus est, lorsque ceux-ci foisonnent. C'est pourquoi il est particulièrement important de toujours replacer cette étape, plus technique certes, au sein du projet global, qui seul lui donne sens : les hypothèses de recherche fournissent le fil directeur de l'analyse des données, filtrent leur exploration, pilotent le projet. Le pourquoi de la manipulation des données ne doit donc jamais être perdu de vue et des échanges réguliers avec l'expert du domaine s'imposent.
Comme nous l'avons vu, des précautions complémentaires sont également nécessaires :
- La phase de constitution du corpus doit sans cesse être questionnée dans la mesure où « l'analyse ne vaudra que ce que vaut le corpus » [94, DALBERA]. L'élaboration du corpus JO a ainsi dû s'adapter à de nombreuses contraintes de sélection des données (disponibilité et lisibilité des données, homogénéité des données, etc.) qui en ont réduit la couverture temporelle. Cette sélection répondait également à des besoins scientifiques, fonction de la problématique (sélection par mots-clés mémoriels, par type d'annonce). L'étape de constitution d'un corpus est donc toujours délicate et de nombreux pièges en jalonnent le parcours. Ainsi, nous avons vu que le filtrage initial des déclarations au moyen du terme générique « m?mo » (pour « mémoire et « commémoration ») a justifié le choix de l'année 2000, alors que la distinction des filtres « mémoire » et « commémoration » désigne plutôt l'année 2001 comme étant celle où le nombre de déclarations contenant « mémoire » dépasse celui des déclarations contenant « anciens combattants ». Ce constat sera pris en compte lors de la poursuite du travail sur les données du JO.
- Un retour systématique aux textes d'origine doit être pratiqué de manière à limiter les éventuels dérapages interprétatifs.
- La prise en compte d'éléments considérés comme peu pertinents d'un point de vue statistique ou d'éléments surprenants s'avère également utile à l'interprétation.
- Enfin, une documentation détaillée de l'ensemble du processus doit avoir lieu afin de pouvoir procéder à des modifications éventuelles ou de permettre une exploitation secondaire des données.
Traiter et interpréter des données n'est jamais un long fleuve tranquille, qui se déroulerait de manière linéaire et extérieure au questionnement scientifique. Sans cesse, des questions et des obstacles imprévus surgissent, auxquels il faut imaginer, collectivement, des alternatives qui ne remettent en question ni la finalité ni l'intégrité du projet. Nous espérons que les éléments exposés tout au long de ce mémoire constituent des exemples convaincants de cette réflexivité propre à toute démarche d'analyse de données textuelles.

Par ailleurs, nous tenons ici à insister sur un aspect peu souvent ou insuffisamment mis en avant : le coût d'apprentissage et le temps que nécessite ce genre de démarche. Il est possible, et même probable (étant donné le caractère très récent de notre investissement dans ce domaine) que la présentation que nous avons faite de la méthode et des résultats obtenus se révèle parfois peu judicieuse voire erronée. Cela nous permet de souligner cet élément primordial : le temps d'apprentissage et d'assimilation de ces technologies est long. Cela implique non seulement la maîtrise de leurs fonctionnalités (loin d'être évidente) mais requiert surtout, et avant tout, la compréhension et l'assimilation des méthodes sous-jacentes, notamment statistiques qui, comme évoquées plus haut, ne sont jamais neutres. Cela demande de tester les outils, de varier les paramètres, de faire le lien entre postulats,

paramètres et résultats mais aussi de lire la littérature sur le sujet, de questionner d'autres utilisateurs plus experts, voire les concepteurs des outils, etc. Autrement dit, comme pour toute technologie un tant soit peu élaborée, mêlant démarches qualitative et quantitative, la simple maîtrise technique, la seule prise en main du logiciel ne sauraient suffire. Et, très certainement, la durée limitée du stage ne nous a pas permis de maîtriser ces principes, de faire un lien plus approfondi entre exposés théoriques et applications concrètes. La conscience de dangers ne nous a pas empêché d'y succomber parfois ! Ainsi, la clarté et la distinction des classes proposées par Alceste ont certainement limité notre étude de la polysémie des termes et du mélange des types de discours. Notre apprentissage ne fait que débuter !
Utiliser ce type d'outils requiert par ailleurs une réelle rigueur. Lors de la manipulation de données, les opérations peuvent vite devenir innombrables, réitérées de multiples fois, parfois pour opérer de mineurs changements – et il est très facile de perdre le fil ! Les versions de fichiers, les traitements effectués se succèdent, tel fichier contenant une information, tel autre l'excluant. Nous avons ainsi vu que les étapes d'élaboration du corpus, de prétraitements et de nettoyage des données étaient constituées d'un ensemble de micro-tâches, chronophages – chacune devant être discutée, analysée. Intégrer une démarche d'analyse de données textuelles au sein d'un projet de recherche ne se réduit donc pas à utiliser un « presse-boutons »… L'influence de ces multiples choix sur les résultats explique qu'une vigilance accrue doit être portée à chaque manipulation. La documentation préconisée du processus prend ici tout son sens.

Perspectives

L'analyse des résultats obtenus avec Calliope et Alceste a permis de proposer quelques pistes de réflexion dans l'optique du prolongement du projet de recherche sur la question mémorielle. Avant de présenter ces pistes, rappelons néanmoins les principaux résultats auxquels nous sommes parvenu, en lien aux hypothèses initiales :
1. Appréhension quantitative et chronologique erronée du phénomène mémoriel ;
2. Appréhension erronée des acteurs-clés du phénomène mémoriel.
S'il y a bien une augmentation de l'emploi du vocabulaire mémoriel sur la période concernée (1984, 2000 et 2010), aucune inflation mémorielle n'est constatée. Les expressions « devoir de mémoire » et « lieu de mémoire » sont certes de plus en plus employées mais sans devenir systématiques (de 0 occurrence en 1984, elles apparaissent dans 2,89% des documents en 2000 et 8,91% des documents en 2010).
Mais plus important, l'emploi de « devoir de mémoire » est plutôt le fait d'associations d'anciens combattants (même si cet emploi baisse dans le temps, de 53% en 2000 à 30% en 2010 selon Calliope) et d'acteurs de conflits passés (43% des occurrences sur toute la période, selon Alceste) – et non de communautés fragmentées, aux « revendications mémorielles particularistes », comme le voudrait l'opinion commune. En cela, les premiers résultats semblent confirmer une partie des hypothèses de S. Gensburger. Les autres occurrences de « devoir de mémoire » ou de « lieu de mémoire » concernent des déclarations en lien avec la conservation du patrimoine.
La distribution géographique a montré, par ailleurs, un lien fort entre Paris et le terme « mémoire » (qui relève souvent du domaine culturel et patrimonial). Il faudrait analyser finement les déclarations concernées pour voir si cette particularité s'inscrit dans l'hypothèse de la tension entre administrations centrales (ministère des Anciens combattants et ministère de la Culture) autour de la thématique mémorielle.

Elargissement de la couverture temporelle du corpus

Dans l'optique d'un prolongement du projet, nous procéderons tout d'abord à un élargissement de la couverture temporelle, de manière à pouvoir mieux répondre aux hypothèses de recherche : chronologie erronée du phénomène mémoriel, plus ancienne que généralement admise (ce phénomène remonterait aux années 70 et non à la fin des années 80). La durée du stage n'a en effet permis de traiter qu'une infime partie des données à

notre disposition – seules trois années de déclarations (1984, 2000 et 2010) ont été analysées sur une période qui s'étend de 1945 à nos jours.
De plus, des fichiers de déclarations au JO qui n'étaient pas manipulables au moment de notre stage le sont désormais, grâce au nettoyage effectué par E. Benaïssa. Nous pourrons donc accroître la fenêtre temporelle des données du JO, en effectuant des sondages sur une période plus large (antérieure aux années 80 et prolongée jusqu'en 2014). Nous tenterons de voir, à cette occasion, si la prise en compte d'une temporalité plus longue vérifie, complète, nuance ou contredit les premiers résultats obtenus – et notamment, si l'hypothèse d'un intérêt des anciens combattants pour la mémoire depuis les années 70 transparaît dans les résultats.

Analyse de la distribution par type de contenu et type d'annonce / rôle des dates
La méthode proposée – regrouper plusieurs années de fichiers en « paquets » – devrait également permettre, par la distinction du type de contenu (titre / objet d'une déclaration) et du type d'annonce (création / modification d'une déclaration), une analyse plus fine de la distribution de la question mémorielle.
Nous tenterons alors de répondre aux questions suivantes : La distinction par type d'annonce constitue-t-elle une variable discriminante ? L'emploi de termes mémoriels caractérise-t-il davantage les associations nouvellement créées ? Les associations déjà existantes modifient-elles leur contenu en raison d'un « effet d'intéressement » ?
Les termes mémoriels apparaissent-ils davantage au niveau du titre que de l'objet de la déclaration ? Et dans ce cas, s'agit-il seulement d'un effet d'annonce, d'une simple opération d'adéquation du lexique à l'importance reconnue d'un objet social désormais légitimé, la « mémoire » ? Si les termes mémoriels figurent au sein des objets des déclarations, peut-on constater une modification substantielle de la finalité de l'association ? Ou ces termes ne sont-ils ajoutés qu'à titre accessoire ?
Une attention particulière sera par ailleurs portée aux déclarations contenant le mot-clé « mémoire ». En soumettant ces déclarations à un traitement distinct, puis en éliminant ce mot-clé de l'analyse lexicale, nous tenterons de voir comment les thématiques et les types de discours mémoriels se déclinent : voir ce qui relève globalement de cette thématique pour ensuite procéder à l'analyse plus fine du vocabulaire : il s'agira de repérer les éventuels glissements et chevauchements lexicaux des déclarations sans que le poids du mot-clé prévale sur les résultats.
Enfin, nous questionnerons le rôle des dates au sein des déclarations : La mention de dates dans le contenu des déclarations constitue-t-elle un nouveau canon mémoriel ? Ces dates sont-elles le signe d'une certaine manière de dire la mémoire ? Sont-elles devenues un passage obligé de la mémoire ?

Multiplication des outils d'analyse de données textuelles
Notre travail nous a par ailleurs convaincu de l'intérêt de multiplier les points de vue sur un même corpus, et donc de varier les outils lexicométriques qui lui sont appliqués. C'est pourquoi l'une des pistes que nous proposons consiste à explorer d'autres outils, fondés sur d'autres méthodes, toujours dans l'optique d'une aide au processus interprétatif.
Nous suggérons d'utiliser Lexico, un outil lexicométrique développé initialement au laboratoire Lexicométrie et textes politiques de l'ENS Fontenay-Saint-Cloud. Cet outil semble particulièrement adapté à nos besoins. Il propose en effet des analyses chronologiques de lexiques par la mise en évidence, au moyen de partitionnement du corpus en périodes, de variations de l'emploi du vocabulaire au fil du temps. Selon [32, GARNIER et GUERIN-PACE], « ses graphiques permettent de représenter très efficacement l'évolution de l'usage d'un ou plusieurs mots dans le temps (cf. l'exemple d'analyse de textes syndicaux dans Lamalle C., Salem A., "Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels", Actes JADT'2002, Journées d'analyse statistique des données textuelles, 2002). »
L'utilisation de ce logiciel pourrait ainsi permettre de comparer ses résultats à ceux obtenus avec Calliope : voir si des différences émergent entre résultats, des complémentarités ou des contradictions qu'il faudra alors tenter de comprendre.
Mais l'utilisation de plusieurs outils a une autre vertu, pédagogique cette fois-ci : outre la complémentarité de perspectives, l'alternance d'outils permet de mieux saisir, lors de

l'apprentissage de logiciels d'analyse textuelle, les fonctionnements de l'un via ses différences avec un autre. La multiplicité d'usage devient ainsi un véritable usage contrastif, à forte vertu pédagogique.

Rôle des ingénieurs au sein de projets d'humanités numériques ?

Nous souhaitons conclure ce travail par une mise en perspective, non plus de la méthode utilisée, mais de la distribution des rôles au sein d'un projet d'humanités numériques. La division du travail que pratique et revendique ce type de démarche facilite la mise en œuvre de projets nécessitant des compétences variées. Nous avons vu, au sein même du projet auquel nous collaborons, que de multiples savoir-faire étaient requis : le pilotage scientifique, bien sûr, mais aussi des compétences en analyse de données textuelles, en langages de programmation (développement de scripts), en méthodes quantitatives, en analyse prosopographique, etc. Cette distribution des tâches entre chercheurs et ingénieurs semble s'imposer en raison, déjà, de la taille des corpus de données à traiter mais aussi de la variété des compétences à réunir.

Nous ne reviendrons pas ici sur l'aspect quantitatif des données (« Big data : is bigger better ? ») mais sur les conséquences de cette division du travail sur le processus scientifique et la place que peut occuper un ingénieur au sein de ce type de projet.

Deux risques se profilent selon nous. Le premier est lié à la dissociation que cette répartition peut introduire entre le moment de l'analyse des données au moyen des outils lexicométriques et celui de l'interprétation. Les biais inhérents à la démarche d'analyse de données textuelles peuvent alors redoubler dans une situation où le chercheur n'est pas aux commandes des outils. Que traiter ? Comment le traiter ? Quels résultats sélectionner ? C'est la raison pour laquelle l'ingénieur qui opère « extérieurement » doit sans cesse dialoguer avec le chercheur : les résultats doivent être discutés collectivement, les questionnements partagés, la sélection des résultats justifiée, la problématique constamment clarifiée, les hypothèses redéfinies et expliquées, etc. Ces échanges visent à rapprocher deux moments distincts d'un même projet, en vue d'une même finalité. La validité de la démarche en dépend.

La seconde limite que peut représenter cette forme d'externalisation d'une partie du processus de recherche consiste à se demander si, d'un point de vue professionnel, cela ne représente pas une spécialisation « forcée » de l'ingénieur. Sélectionner les éléments pertinents à l'interprétation nécessite en effet de comprendre et d'assimiler les hypothèses du projet auquel collabore l'ingénieur. Il ne s'agit pas d'apporter des compétences qui seraient interchangeables mais de s'approprier les enjeux de la recherche pour pouvoir adapter l'analyse des données aux besoins. Une telle spécialisation signifie-t-elle, à terme, une réduction des possibles pour l'ingénieur ? Ou constitue-t-elle une occasion de collaborer plus activement et plus utilement aux projets de recherche ?

Bibliographie

La bibliographie qui suit a été arrêtée le 6 novembre 2015. Tous les liens d'accès aux documents référencés étaient valides à cette date. L'encyclopédie en ligne Wikipédia, qui n'est pas citée dans cette bibliographie, a également été régulièrement utilisée : les liens vers les pages consultées sont indiqués dans le corps du mémoire, en note de bas de page.

Les références bibliographiques sont classées thématiquement, puis par ordre alphabétique d'auteur au sein de chaque thématique. Une numérotation leur est affectée, qui correspond à celle mentionnée dans le corps du mémoire sous la forme [n°, NOM de l'auteur].

LE PHÉNOMÈNE MÉMORIEL

Publications traitant des politiques publiques de la mémoire et la sociologie de la mémoire.

[1] **GENSBURGER S**., *Les Justes de France. Politiques publiques de la mémoire*, Presses de Sciences Po., Paris, 2010

[2] **GENSBURGER S.**, « Comprendre la multiplication des" journées de commémoration nationale": étude d'un instrument d'action publique de nature symbolique », C. HALPERN, P. LASCOUMES et P. LE GALES (dir.), *L'instrumentation de l'action publique. Controverses, résistances, effets*, Presses de Sciences Po, Paris, 2014, p. 345-365

[3] **GENSBURGER S.**, « Mémoire et bricolage », *Ethnologie française*, vol. 37, n°3, 2007, p. 433-440

[4] **GENSBURGER S.**, « Réflexions autour de la "politique de la mémoire". L'exemple de l'évocation des "Justes parmi les Nations" en France », M. OFFERLE, H. ROUSSO (dir.), *La fabrique interdisciplinaire. Histoire et science politique*, Presses universitaires de Rennes, 2008, p. 133-147

[5] **GENSBURGER S. et M.-C. LAVABRE**, « Entre " devoir de mémoire" et "abus de mémoire" : la sociologie de la mémoire comme tierce position », B. MÜLLER (dir.), *Histoire, mémoire et épistémologie. A propos de Paul Ricoeur*, Payot, Lausanne, 2005, p. 76-95

HUMANITÉS NUMÉRIQUES

Publications traitant du mouvement des humanités numériques et ses enjeux (méthodologiques, scientifiques, institutionnels, sociaux, politiques, etc.).

[6] **BERRA A.**, « Faire des humanités numériques », P. MOUNIER (dir.), *Read/Write Book 2*, OpenEdition Press, Paris, 2012

[7] **BOULLIER D. et A. LOHARD**, *Opinion mining et Sentiment analysis*, OpenEdition Press, coll.« Sciences po / Médialab », Paris, 2012

[8] **BOYD D. et K. CRAWFORD**, « Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon », *Information, Communication & Society*, vol. 15, n°5, 2012, p. 662-679 [En ligne : http://www.tandfonline.com/doi/pdf/10.1080/1369118X.2012.678878]

[9] **CECIRE N.**, « When Digital Humanities Was in Vogue », *Journal of Digital Humanities*, 2011, vol. 1, n°1 [En ligne : http://journalofdigitalhumanities.org/1-1/when-digital-humanities-was-in-vogue-by-natalia-cecire]

[10] **CECIRE N.**, « » Introduction: Theory and the Virtues of Digital Humanities *», Journal of Digital Humanities*, 2011, vol. 1, n°1 [En ligne : http://journalofdigitalhumanities.org/1-1/introduction-theory-and-the-virtues-of-digital-humanities-by-natalia-cecire]

[11] **CLAVERT F.**, « Le code et l'historien contemporanéiste – pensées éparses », *L'histoire contemporaine à l'ère numérique*, 2011 [En ligne : http://histnum.hypotheses.org/385]

[12] **CLIVAZ C.**, « "Humanités Digitales" : mais oui, un néologisme consciemment choisi ! Suite », *Digital Humanities Blog*, 2012 (modifié 2013) [En ligne : http://claireclivaz.hypotheses.org/114]

[13] **DACOS M.**, « La stratégie du Sauna finlandais », *Blogo-Numéricus*, 2013 [En ligne : http://bn.hypotheses.org/11138]

[14] **GEFEN A.**, « Les enjeux épistémologiques des humanités numériques », *Socio*, 2015, n°4 [En ligne : https://socio.revues.org/1296]

[15] **LEON J.**, « Le CNRS et les débuts de la traduction automatique en France », *La revue pour l'histoire du CNRS*, 2002, n°6 [En ligne : http://histoire-cnrs.revues.org/3461]

[16] **LIU A.**, « The state of the digital humanities. A report and a critique », *Arts and Humanities in Higher Education* - Special issue, 2011, vol. 11, n°1-2, p. 8-41

[17] **MORETTI F.**, *Graphes, cartes et arbres. Modèles abstraits pour une autre histoire de la littérature*, Les prairies ordinaires, Paris, 2008 (traduction française)

[18] **MOUNIER P.**, « Une "utopie politique" pour les humanités numériques ? Modèles de communication savante et de gestion de la recherche en transformation », *Socio*, 2015, n°4, p. 97-112 [En ligne : https://socio.revues.org/1451]

[19] **MOUNIER P.** (dir.), *Read/Write Book 2*, OpenEdition Press, Paris, 2012

[20] **MOUNIER P.**, « Ce que sait la main », *Blogo-Numericus*, 2010 [En ligne : http://bn.hypotheses.org/10434]

[21] **RUIZ E.**, « Les historiens seront-ils finalement programmeurs ? », *La boîte à outils des historiens*, 2011 [En ligne : http://www.boiteaoutils.info/2011/09/les-historiens-seront-ils-finalement]

[22] **SCHREIBMAN S., R. SIEMENS et J. UNSWORTH** (dir.), *A Companion to Digital Humanities*, Oxford, Blackwell, 2004 [En ligne : http://www.digitalhumanities.org/companion]

[23] **TURKEL W.J. et A. MACEACHERN**, *The Programming Historian*, NiCHE - Network in Canadian History & Environment, Ontario, Canada, 1st Edition, 2007 [En ligne : http://niche-canada.org/wp-content/uploads/2013/09/programming-historian-1.pdf]

[24] **URBANIAK A.**, « Humanités numériques : une question de lexique », *ThatCamp Saint-Malo 2013*, Paris, Éditions de la Maison des sciences de l'homme, 2013 [En ligne : http://books.openedition.org/editionsmsh/2189]

[25] **YVON F.**, « Une petite introduction au traitement automatique des langues naturelles », 2007 [En ligne : http://perso.limsi.fr/Individu/anne/coursM2R/intro.pdf]

ANALYSE DE DONNÉES TEXTUELLES

Publications traitant des différentes méthodes d'analyse de données textuelles et des outils.

[26] **BASSAC C., J. BUSQUETS et M. VERSEL**, « Analyse statistique des données textuelles à partir de publications de Calvet concernant les langues minoritaires », *Lengas*, n°66, 2009, p. 57-78 [En ligne : http://www.msha.fr/baseclme/documents/analyse%20text_Lengas.pdf]

[27] **BEAUDOUIN V.**, « Statistique textuelle : une approche empirique du sens à base d'analyse distributionnelle », *Texto*, 2000 [En ligne : http://www.revue-texto.net/Inedits/Beaudouin_Statistique.html]

[28] **BRUGIDOU M., C. ESCOFFIER, H. FOLCH, S. LAHLOU, D. LE ROUX, P. MORIN-ANDREANI, et G. PIAT**, « Les facteurs de choix et d'utilisation de logiciels d'Analyse de Données Textuelles », 5es Journées internationales d'analyse statistique des données textuelles (JADT), Lausanne, 2000 [En ligne : http://lexicometrica.univ-paris3.fr/jadt/jadt2000/pdf/04/04.pdf]

[29] **COURTIAL J.-P. et L. KERNEUR**, « La méthode des mots associés, outil d'analyse du changement social », *Histoire & Mesure*, 1997, vol. 12, n°3/4, p. 251-270 [En ligne : http://www.persee.fr/doc/hism_0982-1783_1997_num_12_3_1546]

[30] **DELAVIGNE V.**, « Alceste, un logiciel d'analyse textuelle », *Texto, Sémantique des textes*, 2003 [En ligne : https://hal.archives-ouvertes.fr/hal-00924168/document]

[31] **DEMAZIERE D., C. BROSSAUD, P. TRABAL, et K.M. METER (VAN)** (dir.), *Analyses textuelles en sociologie Logiciels, méthodes, usages*, Presses Universitaires de Rennes, 2006

[32] **DUMONT V.**, « Du débat sur la place des logiciels dans l'analyse de données qualitatives », *Recherches qualitatives*, 2010, 9 - Hors-série, p. 1-14 [En ligne : http://www.recherche-qualitative.qc.ca/documents/files/revue/hors_serie/hors_serie_v9/HS9_Intro_Dumont.pdf]

[33] **FALLERY B. et F. RODHAIN**, « Quatre approches pour l'analyse de données textuelles : lexicale, linguistique, cognitive, thématique », 16e Conférence internationale de management stratégique (AIMS), Montréal, juin 2007 [En ligne : https://hal.archives-ouvertes.fr/hal-00821448]

[34] **GARNIER B. et F. GUERIN-PACE**, *Appliquer les méthodes de la statistique textuelle*, CEPED, Paris, 2010 [En ligne : http://www.ceped.org/IMG/pdf/appliquer_les_methodes_de_la_statistique_textuelle-.pdf]

[35] **GEKA M. et M. DARGENTAS**, « L'apport du logiciel Alceste à l'analyse des représentations sociales : l'exemple de deux études diachroniques », *Les Cahiers internationaux de psychologie sociale*, 2010 [En ligne : www.cairn.info/revue-les-cahiers-internationaux-de-psychologie-sociale-2010-1-page-111.htm]

[36] **GUERIN-PACE F.**, « La statistique textuelle. Un outil exploratoire en sciences sociales », *Population*, 1997, vol. 52, n°4, p. 865-887 [En ligne : http://www.cairn.info/revue-population-1997-4-page-865.htm]

[37] **GUILHAUMOU J.**, « L'historien du discours et la lexicométrie Étude d'une série chronologique : le « Père Duchesne » d'Hébert (Juillet 1793 - mars 1794) », *Histoire & Mesure*, 1986, vol. 1, n°3, p. 27-46 [En ligne : http://www.persee.fr/doc/hism_0982-1783_1986_num_1_3_1529]

[38] **JENNY J.**, « A propos des logiciels d'analyse textuelle pratiqués en France pour la recherche en sciences sociales », Legs-sociologique, http://jacquesjenny.com/legs-sociologique/?page_id=1889, 1997

[39] **JENNY J.**, « Méthodes et pratiques formalisées d'analyse de contenu et de discours dans la recherche sociologique française contemporaine. Etat des lieux et essai de classification », *Bulletin de méthodologie sociologique* (BMS), 1997, n°54, p. 64-112 [En ligne : http://jacquesjenny.com/legs-sociologique/?page_id=1253]

[40] **KALAMPALIKIS N.**, « L'apport de la méthode Alceste dans l'analyse des représentations sociales », *Méthodes d'étude des représentations sociales*, Eres, Paris, 2005

[41] **KOMIS V., C. DEPOVER, et T. KARSENTI**, « L'usage des outils informatiques en analyse des données qualitatives », *Adjectif*, 2013 [En ligne : http://www.adjectif.net/spip/spip.php?article216]

[42] **LABBE C. et D. LABBE**, « Lexicométrie : quels outils pour les sciences humaines et sociales ? », Journée d'études « Usages de la lexicométrie en sociologie », juin 2013, Guyancourt [En ligne : https://hal.archives-ouvertes.fr/hal-00834039/document]

[43] **LANDRY N., S. BHANJI-PITMAN, et R. AUGER**, « L'instrumentation dans la collecte des données. Comparaison d'un mode de sélection par le chercheur et d'un mode d'extraction automatisée de données textuelles », *Recherches qualitatives*, 2005, Hors-série, n°2, p. 70-85 [En ligne : http://www.recherche-qualitative.qc.ca/documents/files/revue/hors_serie/hors_serie_v2/NLandry%20et%20al%20HS2-issn.pdf]

[44] **LAROSE F., T. KARSENTI, et V. GRENON**, « Regards sur diverses approches de traitement des données textuelles. Les outils, leurs fondements et l'épistémologie de leurs usages », *Formation et profession*, 2000, n°6/2, p. 5-12 [En ligne : http://www.researchgate.net/publication/267035349_Regards_sur_diverses_approches_de_traitement_des_donnes_textuelles._Les_outils_leurs_fondements_et_l'pistmologie_de_leurs_usages]

[45] **LEJEUNE C.**, « Montrer, calculer, explorer, analyser. Ce que l'informatique fait (faire) à l'analyse qualitative », *Recherches qualitatives*, 2010, n°9, p. 15-32 [En ligne : http://orbi.ulg.ac.be/handle/2268/61098]

[46] **LEJEUNE C.**, « Représentations des réseaux de mots associés », 7es Journées internationales d'analyse statistique de données textuelles (JATD), 2004 [En ligne : http://lexicometrica.univ-paris3.fr/jadt/jadt2004/pdf/JADT_069.pdf]

[47] **LEJEUNE C. et A. BENEL**, « Lexicométrie pour l'analyse qualitative. Pourquoi et comment résoudre le paradoxe ? », 11es Journées internationales d'analyse statistique des données textuelles (JADT), 2012 [En ligne : http://orbi.ulg.ac.be/handle/2268/125414]

[48] **MAYAFFRE D.**, « L'analyse du discours assistée par ordinateur », 2009 [En ligne : http://eprints.aidenligne-francais-universite.auf.org/19]

[49] **MAYAFFRE D.**, « De la lexicométrie à la logométrie », *Astrolabe*, 2005, p. 1-11 [En ligne : https://hal.archives-ouvertes.fr/hal-00551921]

[50] **MAYAFFRE D.**, « Plaidoyer en faveur de l'Analyse de Données co (n) Textuelles. Parcours cooccurrentiels dans le discours présidentiel français (1958-2014) », 12es Journées internationales d'analyse statistique de données textuelles (JADT), Paris, 2014 [En ligne : http://lexicometrica.univ-paris3.fr/jadt/jadt2014/01-ACTES/01-JADT2014.pdf]

[51] **MAYAFFRE D.**, « L'entrelacement lexical des textes. Cooccurrences et lexicométrie », Journées de Linguistique de Corpus, Lorient, 2007 [En ligne : https://hal.archives-ouvertes.fr/hal-00553808]

[52] **MAYAFFRE D.**, « Vers une herméneutique matérielle numérique. Corpus textuels, Logométrie et Langage politique », Mémoire de synthèse en vue de l'Habilitation à diriger des recherches, Université Nice Sophia Antipolis, 2010 [En ligne : https://tel.archives-ouvertes.fr/tel-00655380]

[53] **MARPSAT M.**, « La méthode Alceste », *Sociologi*e, 2010, vol. 1, n°1 [En ligne : https://sociologie.revues.org/312]

[54] **OLLIVIER G.**, « Panorama critique des analyses textuelles informatisées en SHS », Academia.edu, 2009 [En ligne : http://www.academia.edu/2854000/Panorama_critique_des_analyses_textuelles_informatises_en_SHS]

[55] **POLO DE BEAULIEU M.-A.**, « Panorama de la lexicométrie », *Histoire & Mesure*, 1987, II, n°3/4, p. 173-197 [En ligne : http://www.persee.fr/doc/hism_0982-1783_1987_num_2_3_1330]

[56] **QUATRAIN Y., S. NUGIER, A. PERADOTTO, et D. GARROUSTE**, « Evaluation d'outils de Text Mining : démarches et résultats », 7es journées internationales d'analyse statistique des données textuelles (JADT), 2004 [En ligne : http://archivesic.ccsd.cnrs.fr/sic_00001256]

[57] **RATINAUD P.**, « Outils informatiques appliqués aux sciences de l'éducation », support de cours, Université de Toulouse, 2011 [En ligne : http://repere.no-ip.org/Members/pratinaud/informatique/cours_ED355X.pdf]

[58] **RATINAUD P. et P. MARCHAND**, « Application de la méthode ALCESTE à de " gros" corpus et stabilité des "mondes lexicaux " : analyse du "CableGate" avec IRaMuTeQ », 11es Journées internationales d'analyse statistique des données textuelles (JADT), 2012 [En ligne : http://lexicometrica.univ-paris3.fr/jadt/jadt2012/Communications/Ratinaud,%20Pierre%20et%20al.%20-%20Application%20de%20la%20methode%20Alceste.pdf]

[59] **REINERT M.**, « Postures énonciatives et mondes lexicaux stabilisés en analyse statistique de discours », *Langage et société*, 2007, n°121-122, p. 189-202 [En ligne : https://www.cairn.info/revue-langage-et-societe-2007-3-page-189.htm]

[60] **REINERT M.**, « La tresse du sens et la méthode "Alceste". Application aux "Rêveries du promeneur solitaire" », 5es Journées internationales d'analyse statistique des données textuelles (JADT), 2000 [En ligne : http://lexicometrica.univ-paris3.fr/jadt/jadt2000/pdf/31/31.pdf]

[61] **REINERT M.**, « Quelques interrogations à propos de l'"objet" d'une analyse de discours de type statistique et de la réponse "Alceste" », *Langage et société*, 1999, vol. 90, n°1, p. 57-70 [En ligne : http://www.persee.fr/doc/lsoc_0181-4095_1999_num_90_1_2897]

[62] **REINERT M.**, « Quel objet pour une analyse statistiques du discours ? Quelques réflexions à propose de la réponse Alceste », 4es Journées internationales d'analyse statistique des données textuelles (JADT), 1998 [En ligne : http://lexicometrica.univ-paris3.fr/jadt/jadt1998/reinert.htm]

[63] **REINERT M.**, « Alceste une méthodologie d'analyse des données textuelles et une application : Aurélia de Gérard de Nerval », *Bulletin de méthodologie sociologique* (BMS), 1990, vol. 26, n°1, p. 24-54 [En ligne : http://bms.sagepub.com/content/26/1/24.abstract]

[64] **SAINT LEGER (DE) M. et K.M. METER (VAN)**, « L'apport de la scientométrie et la méthode des mots associés », *Analyses textuelles en sociologie Logiciels, méthodes et usages*, Presses universitaires de Rennes, 2006

[65] **SAINT LEGER (DE) M. et K.M. METER (VAN)**, « Cartographie du premier congrès de l'ASF avec la méthode des mots associés », *Bulletin de méthodologie sociologique* (BMS), 2005 [En ligne : http://bms.revues.org/1050]

[66] **SAINT LEGER (DE) M.**, « Comment ont évolué les thématiques des 99 premiers numéros de BMS ? Analyse avec un logiciel de fouille de texte », *Bulletin de méthodologie sociologique* (BMS), 2008 [En ligne : http://bms.revues.org/3153]

[67] **SAINT LEGER (DE) M.**, *Modélisation des flux d'information scientifique et technique par le bruit : Vers un suivi des domaines de la connaissance*, Thèse de doctorat, CNAM, 1997

MÉTHODES QUALITATIVES ET QUANTITATIVES

Publications traitant des méthodes quantitatives et qualitatives, des principes statistiques et linguistiques sur lesquels se fonde l'analyse de données textuelles.

[68] **BLONDIAUX L.**, « Le chiffre et la croyance. L'importation des sondages d'opinion en France ou les infortunes d'une opinion sans publics », *Politix*, 1994, vol. 7, n°25, p. 117-152 [En ligne : http://www.persee.fr/doc/polix_0295-2319_1994_num_7_25_1828]

[69] **BRUNET E.**, « Le lemme comme on l'aime », 6es Journées internationales d'analyse statistique des données textuelles (JADT), 2002 [En ligne : http://lexicometrica.univ-paris3.fr/jadt/jadt2002/PDF-2002/brunet.pdf]

[70] **CAREY S.**, « Data is data, or are they? », Sentence first, 2009 [En ligne : https://stancarey.wordpress.com/2009/05/07/data-is-data-or-are-they]

[71] **CHARPENTIER A.**, « Donner », Freakonometrics, 2013 [En ligne : http://freakonometrics.hypotheses.org/11037]

[72] **CHARPENTIER A.**, « Raw Data » is an Oxymoron, Freakonometrics, 2014 [En ligne : http://freakonometrics.hypotheses.org/16777]

[73] **CORMAN J.**, « Extraction d'expressions polylexicales sur corpus arboré », *Linguistics*, 2012 [En ligne : http://dumas.ccsd.cnrs.fr/dumas-00704873/document]

[74] **DANTIER B.**, « "La représentation et l'étude visuelles des informations" », [Extrait de] *Sémiologie graphique*, Paris, Editions EHESS, 2008

[75] **DESROSIERES A.**, *La politique des grands nombres. Histoire de la raison statistique*, Paris, Éditions La Découverte, 2010

[76] **DUCHESNE S.**, « Développement de l'analyse secondaire et des méthodes d'analyse qualitative : une chance à saisir? », M. BRUGIDOU, M. DARGENTAS, D. LE ROUX et A.C. SALOMON (dir.), *Analyse secondaire en recherche qualitative : enjeux pour les sciences*

humaines et sociales, Editions Lavoisier, 2007 [En ligne : https://halshs.archives-ouvertes.fr/halshs-00841876]

[77] **DUFOUR D.** (dir.), *Images à Charge. La construction de la preuve par l'image*, Editions Xavier Barral - Le Bal, Paris, 2015

[78] **GRANDJEAN M.**, « Introduction à la visualisation de données : analyse de réseau en histoire », *Geschichte und Informatik*, 2015, n°18/19, p. 109-128 [En ligne : http://www.martingrandjean.ch/wp-content/uploads/2015/09/Grandjean2015.pdf]

[79] **JENNY J.**, « "Quanti / Quali", Distinction fallacieuse et stérile ! », 1er Congrès de l'Association française de sociologie (AFS), Villetaneuse, 2004 [En ligne : http://jacquesjenny.com/legs-sociologique/?page_id=1159]

[80] **JENNY J.**, « Pour engager un débat avec Max Reinert, à propos des fondements théoriques et des présupposés des logiciels d'analyse textuelle », *Langage et société*, vol. 90, n°1, p. 73-85 [En ligne : http://www.persee.fr/doc/lsoc_0181-4095_1999_num_90_1_2899]

[81] **LEBART L., M. PIRON, et J.-F. STEINER**, *La sémiométrie : essai de statistique structurale* [En ligne], Dunod, 2003 [En ligne : http://horizon.documentation.ird.fr/exl-doc/pleins_textes/divers11-01/010035318.pdf]

[82] **LEBART L. et A. SALEM**, *Statistique textuelle,* Dunod, Paris, 1994

[83] **LEMERCIER C. et C. ZALC**, *Méthodes quantitatives pour l'historien*, Éditions La Découverte, Paris, 2008

[84] **MAYAFFRE D.**, «" Ça suffit comme ça! ". La fausse opposition quantitatif/qualitatif à l'épreuve du discours sarkozyste », *Corela (Cognition, représentation, langage)*, 2014, n° HS-15 [En ligne : http://corela.revues.org/3543]

[85] **PERETTI DE G.**, « La "mise en variables" des textes : mythe ou réalité ? », *Bulletin de méthodologie sociologique* (BMS), 2008, n°88 [En ligne : http://bms.revues.org/773]

[86] **PETIT G.**, « Lemmatisation et figement lexical : les locutions de type SV », *Cahiers de lexicologie*, 2003, vol. 82, n°1, p. 30-57 [En ligne : https://hal.archives-ouvertes.fr/hal-00644420]

[87] **RAJMAN M. et L. LEBART**, « Similarités pour données textuelles », 4es Journées internationales d'analyse statistique des données textuelles (JADT), 1998 [En ligne : http://liawww.epfl.ch/Publications/Archive/RajmanLebart98.pdf]

[88] **REINERT M. et J. JENNY**, « A propos des fondements théoriques et des présupposés des logiciels d'analyse textuelle », *Langage et Société*, 1999, n°90, p. 57-85 [En ligne : http://www.persee.fr/doc/lsoc_0181-4095_1999_num_90_1_2897]

[89] **ROUX M.**, *Algorithmes de classification*, Editions Masson, Paris, 1985 [En ligne : http://www.imep-cnrs.com/mroux/algoclas.pdf]

[90] **SARFATI G.-E.**, *Eléments d'analyse du discours*, Armand Colin., Paris, 1997

[91] **TRUDEL L., C. SIMARD, et N. VONARX**, « La recherche qualitative est-elle nécessairement exploratoire ? », *Recherches qualitatives*, 2007, Hors-série, vol. 5, p. 38-45 [En ligne : http://www.recherche-qualitative.qc.ca/documents/files/revue/hors_serie/hors_serie_v5/trudel.pdf]

MÉTHODES DE CONSTITUTION DE CORPUS, DE PRÉTRAITEMENT ET NETTOYAGE DES DONNÉES

Publications traitant des méthodes de constitution de corpus, de prétraitement et nettoyage des données.

[92] **BOULLIER D. et A. LOHARD**, « Constituer et traiter le corpus de travail, avant l'analyse de tonalité », *Opinion mining et Sentiment analysis*, Paris, OpenEdition Press, 2012

[93] **CHARAUDEAU P.**, « Dis-moi quel est ton corpus, je te dirai quelle est ta problématique », *Corpus*, 2009, n°8, p. 37-66 [En ligne : http://corpus.revues.org/1674]

[94] **DALBERA J.-P.**, « Le corpus entre données, analyse et théorie », *Corpus*, 2002, n°1 [En ligne : http://corpus.revues.org/10]

[95] **DENIS J. et S. GOËTA**, « La fabrique des données brutes. Le travail en coulisses de l'open data », Journées d'études SACRED « Penser l'écosystème des données. Les enjeux scientifiques et politiques des données numériques », Paris, février 2013 [En ligne : https://halshs.archives-ouvertes.fr/halshs-00990771]

[96] **HABERT B.**, « Des corpus représentatifs : de quoi, pour quoi, comment ? », M. BILGER (dir.), *Linguistique sur corpus. Études et réflexions*, Perpignan, Presses Universitaires de Perpignan, 2000, p. 11-58

[97] **HABERT B., A. NAZARENKO et A. SALEM**, *Les linguistiques de corpus*, Armand Colin, Paris, 1997 [En ligne : http://lexicometrica.univ-paris3.fr/livre/les_linguistiques_de_corpus_1997/les_linguistiques_de_corpus_1997.pdf]

[98] **HOOLAND S. van et R. VERBORGH**, *Linked data for libraries, archives and museums: how to clean, link and publish your metadata*, Facet Publishing, London, 2014

[99] **LAHIRE B.**, « Remarques sociologiques sur le *linguistice turn*. Suite au "Dialogue sur l'espace public" entre Keith M. Baker et Roger Chartier », *Politix*, 1994, vol. 7, p. 189-192 [En ligne : http://www.persee.fr/doc/polix_0295-2319_1994_num_7_27_1871]

[100] **LAHIRE B.**, « Précisions sur la manière sociologique de traiter du "sens" : quelques remarques concernant l'ethnométhodologie », *Langage et société*, 1992, n°59, p. 73-89 [En ligne : http://www.persee.fr/doc/lsoc_0181-4095_1992_num_59_1_2560]

[101] **MARSHMAN E.**, « Construction et gestion des corpus : Résumé et essai d'uniformisation du processus pour la terminologie », [En ligne : http://olst.ling.umontreal.ca/pdf/terminotique/corpusentermino.pdf]

[102] **MAYAFFRE D.**, « Les corpus politiques : objet, méthode et contenu », *Corpus*, 2005, n°4, p. 5-19 [En ligne : http://corpus.revues.org/292]

[103] **MAYAFFRE D.**, « Effervescence autour des corpus », M. BALLARD et C. PINEIRA (dir.), *Corpus en linguistique et en traductotologie*, Artois Presses Université, 2007, p. 61-71 [En ligne : https://halshs.archives-ouvertes.fr/hal-00912006/document]

[104] **PERY-WOODLEY M.-P.**, « Quels corpus pour quels traitements automatiques ? », *Traitement Automatique des Langues*, 1995, vol. 1-2, n°36, p. 213-232 [En ligne : http://w3.erss.univ-tlse2.fr:8080/index.jsp?perso=pery&subURL=articles/TALcorpus.pdf]

[105] **PREVOST S.**, « Corpus informatisés de français médiéval : contraintes sur leur constitution et spécificités de leurs apports », *Corpus*, 2008, n°7 [En ligne : http://corpus.revues.org/1500]

[106] **VERBORGH R. et M. DE WILDE**, *Using OpenRefine. The essential OpenRefine guide*, Packt Publishing, 2013

Annexes

Annexe 1
Listes des rapports officiels et des journées nationales de commémoration

Liste des rapports officiels publiés et des journées nationales de commémorations instaurées depuis les années 80 (et les 3 journées de commémoration qui existaient avant 1983)[248] :

1) Cinq rapports officiels sur les politiques publiques et la mémoire en 2008

- Rapport de la Commission "Becker". Pour la commémoration du 90e anniversaire de 1918, présidé par J.J. Becker, 19 décembre 2007.

- Livre blanc sur la Défense et la sécurité nationale, dernière partie sur le "devoir de mémoire", juin 2008, p. 309-312.

- Rapport sur l'enseignement de la Shoah dans le primaire, présidée par H. Waysbord-Loing, remis au ministre de l'Education nationale, juin 2008.

- Rapport de la commission de réflexion sur la modernisation des commémorations publiques, présidée par A. Kaspi, novembre 2008.

- Rapport d'information au nom de la mission d'information sur les questions mémorielles, dit "Rapport Accoyer" du nom de son président B. Accoyer, enregistré à la Présidence de l'Assemblée nationale en novembre 2008.

2) Onze journées nationales de commémoration, instaurées entre 1983 et 2014

- Loi n°2013-642 du 19 juillet 2013 relative à l'instauration du 27 mai comme journée nationale de la Résistance.

- Loi n°2012-1361 du 6 décembre 2012 relative à la reconnaissance du 19 mars comme journée nationale du souvenir et de recueillement à la mémoire des victimes civiles et militaires de la guerre d'Algérie et des combats en Tunisie et au Maroc.

- Loi n°2012-273 du 28 février 2012 fixant au 11 novembre la commémoration de tous les morts pour la France.

- Décret n°2006-388 du 31 mars 2006 fixant la date en France métropolitaine de la commémoration annuelle de l'abolition de l'esclavage, au 10 mai.

- Décret n°2006-313 du 10 mars 2006 instituant le 18 juin de chaque année une journée nationale commémorative de l'appel historique du général de Gaulle à refuser la défaite et à poursuivre le combat contre l'ennemi.

[248] Cette liste a été élaborée par Sarah Gensburger.

- Décret n°2005-547 du 26 mai 2005 instituant une journée nationale d'hommage aux "morts pour la France" en Indochine, le 8 juin de chaque année.

- Loi n°2003-925 du 26 septembre 2003 instituant une journée nationale d'hommage aux "morts pour la France" pendant la guerre d'Algérie et les combats du Maroc et de la Tunisie, le 5 décembre de chaque année.

- Décret du 31 mars 2003 instituant une journée nationale d'hommage aux Harkis et autres membres des formations supplétives.

- Loi n°2000-644 du 10 juillet 2000 instaurant une journée nationale à la mémoire des victimes des crimes racistes et antisémites de l'Etat français et d'hommage aux "Justes" de France (cette loi abrogera le décret n°93-150).

- Décret n°93-150 du 3 février 1993 instituant une journée nationale commémorative des persécutions racistes et antisémites commises sous l'autorité de fait dite "Gouvernement de l'Etat français" (1940-1944).

- Loi n°83-550 du 30 juin 1983 relative à la commémoration de l'abolition de l'esclavage, jour férié dans les départements de la Guadeloupe, de la Guyane, de la Martinique et de la Réunion et dans la collectivité territoriale de Mayotte.

3) Trois journées nationales de commémoration instaurées entre 1914 et 1982

- Loi n°54-415 du 14 avril 1954 consacrant le dernier dimanche d'avril au souvenir des victimes de la déportation et morts dans les camps de concentration du III^{e} Reich au cours de la guerre de 1939-1945.

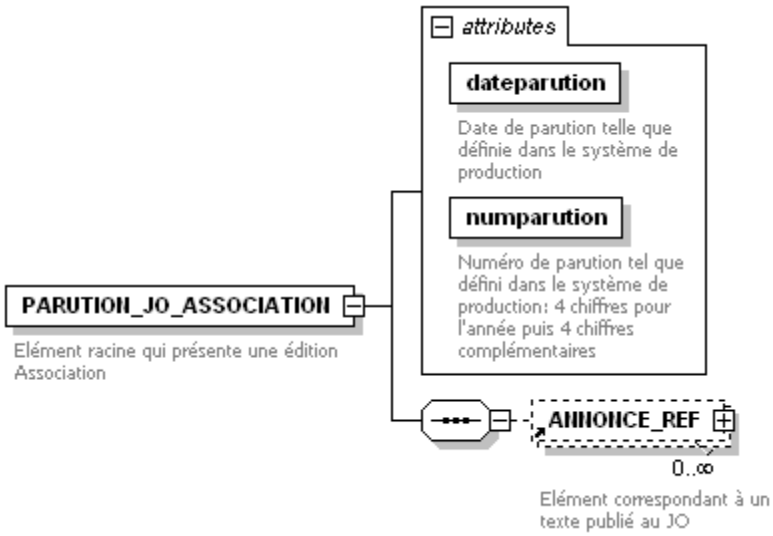
- Loi du 24 octobre 1922 fixant au 11 novembre la commémoration de la victoire et de la paix.

- Loi du 10 juillet 1920 La fête nationale de Jeanne d'Arc et du patriotisme le 2^{e} dimanche de mai.

Annexe 2
Présentation de la DTD[249] du *Journal officiel* (DILA)

1) Présentation de l'élément PARUTION_JO_ASSOCIATION

Chaque publication « association » se qualifie par son numéro de parution et sa date de parution.



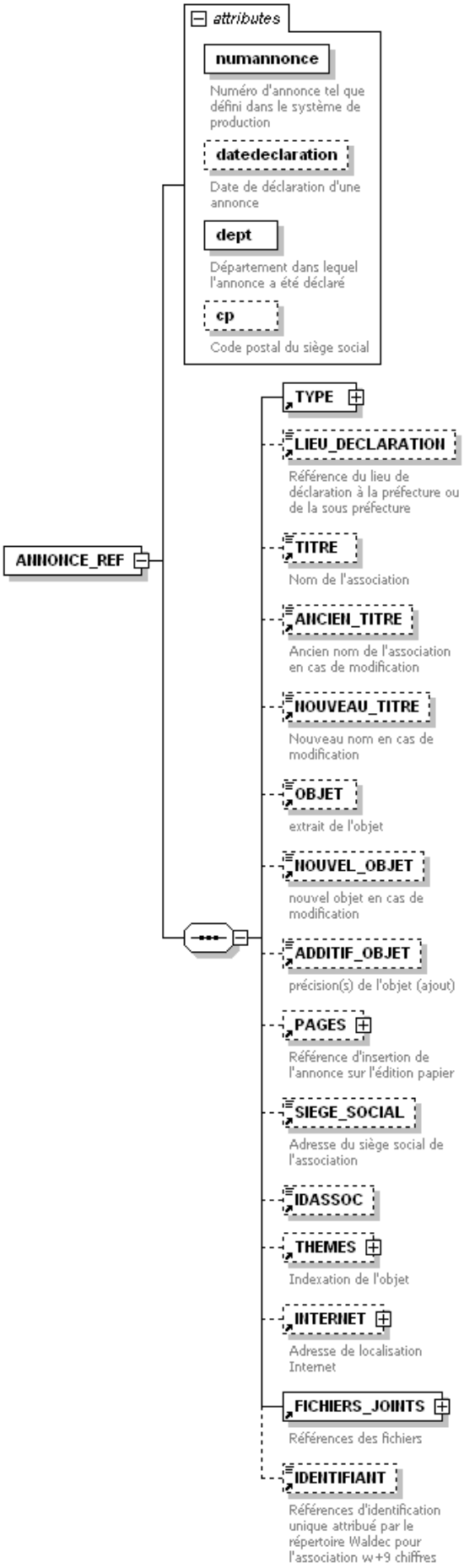
Ex :

<PARUTION_JO_ASSOCIATION dateparution="01/10/2011" numparution="20110040">

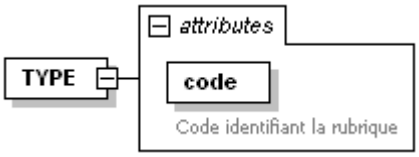
2) Présentation de l'élément ANNONCE_REF

L'élément ANNONCE_REF correspond à la publication d'un avis (le terme d'annonce est utilisé dans la DTD)

[249] Cette présentation de la DTD (Document Type Definition) du JOAFE (*Journal officiel des associations et des fondations d'entreprise*) est issue de la documentation technique publiée par la DILA (direction de l'information légale et administrative), attachée aux services du Premier ministre.


attributes
numannonce
Numéro d'annonce tel que défini dans le système de production
datedeclaration
Date de déclaration d'une annonce
dept
Département dans lequel l'annonce a été déclaré
cp
Code postal du siège social
ANNONCE_REF
TYPE
LIEU_DECLARATION
Référence du lieu de déclaration à la préfecture ou de la sous préfecture
TITRE
Nom de l'association
ANCIEN_TITRE
Ancien nom de l'association en cas de modification
NOUVEAU_TITRE
Nouveau nom en cas de modification
OBJET
extrait de l'objet
NOUVEL_OBJET
nouvel objet en cas de modification
ADDITIF_OBJET
précision(s) de l'objet (ajout)
PAGES
Référence d'insertion de l'annonce sur l'édition papier
SIEGE_SOCIAL
Adresse du siège social de l'association
IDASSOC
THEMES
Indexation de l'objet
INTERNET
Adresse de localisation Internet
FICHIERS_JOINTS
Références des fichiers
IDENTIFIANT
Références d'identification unique attribué par le répertoire Waldec pour l'association w+9 chiffres

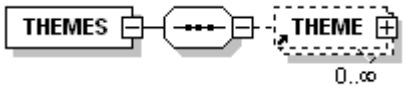
3) Présentation de l'élément TYPE

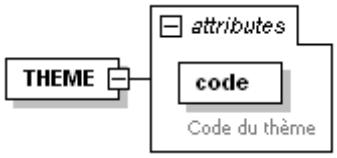


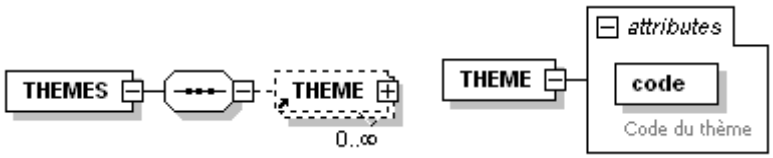
Valeur du Code[250]	Correspond à la Rubrique
1	**Création d'association**
11	Rectificatif de création d'association
111	Annulation de création d'association
2	**Modification d'association**
22	Rectificatif de modification d'association
222	Annulation de modification d'association
3	**Dissolution d'association**
33	Rectificatif de dissolution d'association
333	Annulation de dissolution d'association
4	**Création de fondation d'entreprise**
44	Rectificatif de création de fondation d'entreprise
444	Annulation de création de fondation d'entreprise
5	**Modification de fondation d'entreprise**
55	Rectificatif de modification de fondation d'entreprise
555	Annulation de modification de fondation d'entreprise
6	**Dissolution de fondation d'entreprise**
66	Rectificatif de dissolution de fondation d'entreprise
666	Annulation de dissolution de fondation d'entreprise
8CA	Annulation de création d'association syndicale de propriétaires
8DA	Annulation de dissolution d'association syndicale de propriétaires
8MA	annulation de modification d'association syndicale de propriétaires
8C	Création d'association syndicale de propriétaires
8CR	Rectificatif de création d'association syndicale de propriétaires
8D	Dissolution d'association syndicale de propriétaires
8DR	Rectificatif de dissolution d'association syndicale de propriétaires
8M	Modification d'association syndicale de propriétaires
8MR	Rectificatif de modification d'association syndicale de propriétaires

[250] Seuls les associations (codes de 1 à 3) intéressent notre travail (les autres annonces concernent les fondations d'entreprise, non prises en compte).

4) Présentation de l'élément THEMES







1 Activités politiques
2 Clubs, cercles de réflexion
3 Défense de droits fondamentaux, activités civiques
4 Justice
5 Information communication
6 Culture, pratiques d\'activités artistiques, culturelles
7 Clubs de loisirs, relations
9 Action socio-culturelle
10 Préservation du patrimoine
11 Sports, activités de plein air
13 Chasse pêche
14 Amicales, groupements affinitaires, groupements d'entraide (hors défense de droits fondamentaux)
15 Education formation
16 Recherche
17 Santé
18 Services et établissements médico-sociaux
19 Interventions sociales
20 Associations caritatives, humanitaires, aide au développement, développement du bénévolat
21 Services familiaux, services aux personnes âgées
22 Conduite d\'activités économiques
23 Représentation, promotion et défense d\'intérêts économiques
24 Environnement, cadre de vie
30 Aide à l'emploi, développement local, promotion de solidarités économiques, vie locale
32 Logement
34 Tourisme
36 Sécurité, protection civile
38 Armée (dont préparation militaire, médailles)
40 Activités religieuses, spirituelles ou philosophiques
50 Domaines divers, domaines de nomenclature SITADELE à reclasser

Annexe 3
Nombre de déclarations au JO par an (1963-1984 et 1997-2014)

Nombre total de déclarations annuelles sur les périodes 1963-1984 et 1997-2014 :

Année (1963-1984)	**Nombre de déclarations**	**Année** (1997-2014)	**Nombre de déclarations**
1963	4 800	1997	72 138
1966	9 300	1998	111 593
1969	13 600	1999	107 619
1970	13 800	2000	91 557
1971	18 000	2001	120 020
1972	20 600	2002	101 899
1973	17 500	2003	122 703
1974	18 500	2004	102 091
1975	17 900	2005	108 754
1976	19 000	2006	104 147
1977	20 000	2008	106 811
1978	22 000	2009	105 234
1979	22 000	2010	96 803
1980	24 000	2011	94 982
1981	23 000	2012	94 602
1982	13 000	2013	95 815
1983	27 000	2014	100 471
1984	29 900		
Total	**333 900**	**Total**	**1 737 239**

Annexe 4
Nombre de déclarations au JO par mot-clé (1997-2014)

(Voir page suivante)

			MÉMOIRE / COMM.				CELEBRATION				PATRIMOINE				SOUVENIR				ANC. COMBATTANT						
	Nombre total asso	Nombre total associations sans doublons	Titres	Objets	Titres + Objets	Taux (Titres + Objets)	Titres	Objets	Titres + Objets	Taux	Titres	Objets	Titres + Objets	Taux	Titres	Objets	Titres + Objets	Taux	Titres	Objets	Titres + Objets	Taux	Total annuel (Titres + Objets)	Taux annuel	
1997	119 725	72 138	102	213	275	0,38%	5	92	96	0,13%	263	927	1042	1,44%	31	96	122	0,17%	298	101	321	0,44%	**1856**	**2,57%**	
1998	141 845	111 593	192	362	488	0,44%	14	108	120	0,11%	415	1588	1768	1,58%	45	190	226	0,20%	408	131	454	0,41%	**3056**	**2,74%**	
1999	132 917	107 619	218	402	546	0,51%	13	174	182	0,17%	518	2221	2433	2,26%	59	410	454	0,42%	416	165	488	0,45%	**4103**	**3,81%**	
2000	153 738	91 557	162	478	581	0,63%	9	117	125	0,14%	283	1197	1315	1,44%	43	547	583	0,64%	420	274	613	0,67%	**3217**	**3,51%**	
2001	155 865	120 020	188	647	771	0,64%	18	143	157	0,13%	491	2071	2251	1,88%	40	310	337	0,28%	480	260	641	0,53%	**4157**	**3,46%**	
2002	142 461	101 899	178	555	675	0,66%	9	132	139	0,14%	428	1880	2052	2,01%	48	343	384	0,38%	493	279	618	0,61%	**3868**	**3,80%**	
2003	219 292	122 703	212	636	771	0,63%	5	212	217	0,18%	488	2210	2400	1,96%	61	309	355	0,29%	454	177	521	0,42%	**4264**	**3,48%**	
2004	179 299	102 091	188	547	676	0,66%	12	164	173	0,17%	426	1859	2011	1,97%	32	230	252	0,25%	326	159	411	0,40%	**3523**	**3,45%**	
2005	169 343	108 754	208	618	747	0,69%	4	190	193	0,18%	427	2005	2159	1,99%	36	265	290	0,27%	322	122	384	0,35%	**3773**	**3,47%**	
2006	154 194	104 147	208	549	675	0,65%	6	188	194	0,19%	443	2046	2188	2,10%	48	186	224	0,22%	251	102	298	0,29%	**3579**	**3,44%**	
2008	147 930	106 811	224	625	745	0,70%	6	197	203	0,19%	429	2381	2538	2,38%	52	161	200	0,19%	262	93	305	0,29%	**3991**	**3,74%**	
2009	136 454	105 234	211	655	781	0,74%	3	205	207	0,20%	495	2478	2639	2,51%	46	199	234	0,22%	229	123	294	0,28%	**4155**	**3,95%**	
2010	123 030	96 803	193	594	721	0,74%	6	183	187	0,19%	431	2279	2411	2,49%	30	185	207	0,21%	228	160	324	0,33%	**3850**	**3,98%**	
2011	121 558	94 982	162	585	677	0,71%	6	189	195	0,21%	379	2259	2377	2,50%	26	148	169	0,18%	176	70	212	0,22%	**3630**	**3,82%**	
2012	122 194	94 602	170	550	636	0,67%	6	166	169	0,18%	403	2430	2565	2,71%	36	177	205	0,22%	150	74	197	0,21%	**3772**	**3,99%**	
2013	123 995	95 815	163	558	643	0,67%	6	200	206	0,21%	378	2277	2402	2,51%	42	170	197	0,21%	148	99	210	0,22%	**3658**	**3,82%**	
2014	131 340	100 471	184	641	741	0,74%	5	233	236	0,23%	448	2694	2807	2,79%	33	187	210	0,21%	140	81	187	0,19%	**4181**	**4,16%**	
Total	**2 475 180**	**1 737 239**	**3 163**	**9 215**	**11 149**	**0,64%**	**133**	**2 893**	**2 999**	**0,17%**	**7 145**	**34 802**	**37 358**	**2,15%**	**708**	**4 113**	**4 649**	**0,27%**	**5 201**	**2 470**	**6 478**	**0,37%**	**62 633**	**3,60%**	**Taux annuel moyen** (tous mots clés confondus)
Taux par mot clé						**0,64%**				**0,17%**				**2,15%**				**0,27%**				**0,37%**	**0,72%**	**Taux moyen mots clés**	
														59,65	poids "patrimoine" relativements aux autres mots clés										

Annexe 5
Variantes du contenu des déclarations d'associations d'anciens combattants relevant de fédérations nationales

Relevé (non exhaustif) des variantes de contenu des déclarations d'associations d'anciens combattants, relevant de fédérations nationales (relevé obtenu avec la fonction « Clustering », méthodes Key colllision ou Nearest neighbor d'OpenRefine). Ce décompte concerne la seule année 2000.

221 déclarations (sur 557 déclarations) avec "UNION NATIONALE DES COMBATTANTS" ou "UNC" dans Titre et dont le contenu (Titre ou Objet) varie :

- **56 déclarations** avec "UNION NATIONALE DES COMBATTANTS" et "maintenir les liens de camaraderie, défendre les intérêts des adhérents, perpétuer le souvenir des combattants morts pour la France" dans objet (toutes ces déclarations sont situées dans département 59)
- **73 déclarations** avec "maintenir les liens de camaraderie ; défendre les intérêts des adhérents ; perpétuer le souvenir des combattants morts pour la France" dans Objet (la seule différence avec les déclarations ci-dessus : le « ; » a remplacé la « , ») (72 déclarations dans le département du Nord (59) et 1 déclaration dans le 62).
- **25 déclarations** avec "maintenir, dans l'intérêt du pays, les liens de camaraderie, d'amitié et de solidarité qui existent entre ceux qui ont participé à la défense de la patrie ; défendre les intérêts moraux, sociaux et matériels de ses adhérents et de leurs ayants droit ; perpétuer le souvenir des combattants morts pour la France ; entretenir et développer des relations fraternelles entre les anciens combattants des nations amies ou alliées"
- **7 déclarations** avec "maintenir des liens de camaraderie ; défendre les intérêts des adhérents ; perpétuer le souvenir des combattants morts pour la France"
- **24 déclarations** avec "maintenir les liens de bonne camaraderie créés entre les combattants de toutes les générations ; défendre les intérêts moraux, sociaux et matériels de ces combattants ; perpétuer le souvenir des combattants morts pour la patrie et servir leur mémoire"
- **10 déclarations** avec "maintenir les liens de camaraderie, d'amitié et de solidarité entre ceux qui ont participé à la défense de la patrie ; défendre les intérêts moraux, sociaux et matériels de ses adhérents ; perpétuer le souvenir des combattants morts pour la France"
- **4 déclarations** avec "maintenir les liens de camaraderie, d'amitié et de solidarité qui existent entre tous ceux qui ont participé à la défense de la patrie ; défendre les intérêts moraux, sociaux et matériels de ses adhérents et leurs ayants droit ; perpétuer le souvenir des combattants morts pour la France"
- **2 déclarations** avec "maintenir dans l'intérêt supérieur du pays les liens de camaraderie et de solidarité qui existent entre ceux qui ont participé à la défense de la patrie ; défendre les intérêts moraux, sociaux et matériels de ses adhérents et leurs ayants droit ; perpétuer le souvenir des combattants morts pour la France, servir leur mémoire ; inciter ses membres à participer activement à la vie de la cité"
- **2 déclarations** avec "maintenir, dans l'intérêt supérieur du pays, les liens de camaraderie, d'amitié et de solidarité qui existent entre ceux qui ont participé à la défense de la patrie, et notamment ceux qui ont vocation à relever de l'organisme officiel en charge des anciens combattants et victimes de guerre"

- **3 déclarations** avec "maintenir les liens de camaderie ; défendre les intérêts des adhérents ; perpétuer le souvenir des combattants morts pour la France"
- **2 déclarations** avec "maintenir les liens de camaraderie, d'amitié et de solidarité qui existent entre ceux qui ont participé à la défense de la Patrie, et notamment ceux qui ont vocation à relever de l'organisme officiel en charge des anciens combattants et victimes de guerre ; défendre les intérêts moraux, sociaux et matériels de ses adhérents et leurs ayants droit ; perpétuer le souvenir des combattants morts pour la France, servir leur mémoire"
- **11 déclarations** avec "maintenir des liens de camaraderie, d'amitié et de solidarité entre ceux qui ont participé à la défense de la patrie ; défendre les intérêts moraux, sociaux et matériels de ses adhérents et de leurs ayants droit ; perpétuer le souvenir des combattants morts pour la France"
- **2 déclarations** avec "maintenir dans l'intérêt supérieur du pays les liens de camaraderie, d'amitié et de solidarité qui existent entre ceux qui ont participé à la défense de la Patrie, et notamment ceux qui ont vocation à relever de l'organisme officiel en charge des anciens combattants et victimes de guerre ; défendre par tous moyens en son pouvoir, les intérêts moraux, sociaux et matériels de ses adhérents et de leurs ayants-droit (ascendants, descendants, conjoints, orphelins) ; perpétuer dans la France métropolitaine, dans les départements d'outre-mer et dans le territoires d'outre-mer, comme chez nos alliés, le souvenir des combattants morts pour la France, servir leur mémoire, entretenir et développer des relations fraternelles entre les anciens combattants des nations amies ou alliées"

35 déclarations avec U.N.C. dans Titre et contenu variable de l'Objet :

- comme "maintenir les liens de camaraderie, d'amitié et de solidarité entre tous ceux qui ont participé à la défense de la patrie ; défendre les intérêts moraux et sociaux de ses adhérents ; perpétuer le souvenir des combattants morts pour la France"
- "maintenir les liens de camaraderie, d'amitié et de solidarité ; défendre les droits des anciens combattants",
- "maintenir les liens de camaraderie, d'amitié et de solidarité qui existent entre tous les anciens combattants qui ont participé à la défense de la patrie et préserver leurs droits ; perpétuer le souvenir des combattants morts pour la France"
- "maintenir les liens de camaraderie, d'amitié et de solidarité qui existent entre ceux qui ont participé à la défense de la patrie ; perpétuer le souvenir des combattants morts pour la France"
- "maintenir les liens de camaraderie, d'amitié entre tous ceux qui ont participé à la défense de la patrie ; défendre les intérêts moraux, sociaux et matériels de ses adhérents et de leurs ayants droit ; perpétuer le souvenir des combattants morts pour la France ; servir leur mémoire ; développer des relations fraternelles entre les anciens combattants des nations amies ou alliées"
- "maintenir, dans l'intérêt supérieur du pays, les liens de camaraderie, d'amitié et de solidarité qui existent entre ceux qui ont participé à la défense de la patrie et notamment ceux qui ont vocation à relever de l'organisme officiel en charge des anciens combattants et victimes de guerre"
- "maintenir les liens de camaraderie, d'amitié et de solidarité qui existent entre tous ceux qui ont participé à la défense de la patrie ; défendre les intérêts moraux, sociaux et matériels de ses adhérents et de leurs ayants droit ; perpétuer le souvenir des combattants morts pour la France, de servir leur mémoire et développer des relations fraternelles entre les anciens combattants des nations amies ou alliées en organisant des cérémonies commémoratives,

patriotiques ou religieuses, dont la guerre, la paix, l'humanisme fourniront l'inspiration et en participant aux manifestations de même nature auxquelles l'union ou ses adhérents seraient conviés"

- "maintenir les liens de camaraderie, d'amitié et de solidarité qui existent entre tous ceux qui ont participé à la défense de la patrie, défendre les intérêts moraux, sociaux et matériels de ses adhérents et de leurs ayants droit, perpétuer le souvenir des combattants morts pour la France, développer des relations fraternelles entre les A.C. des nations amies ou alliées"
- "maintenir dans l'intérêt supérieur du pays, les liens de camaraderie, d'amitié et de solidarité qui existent entre ceux qui ont participé à la défense de la patrie, et notamment ceux qui ont vocation à relever de l'organisme officiel en charge des anciens combattants et victimes de guerre ; défendre, par tous les moyens en son pouvoir, les intérêts moraux, sociaux et matériels de ses adhérents et de leurs ayant-droits (ascendants, descendants, conjoints, orphelins) ; perpétuer dans la France métropolitaine, dans les départements d'outre-mer, comme chez nos alliés, le souvenir des combattants morts pour la France, servir leur mémoire, entretenir et développer des relations fraternelles entre les anciens combattants des nations amies ou alliées", etc.

64 déclarations avec "F.N.A.C.A." dans Titre, dont Objet :

- **2 déclarations** avec "entretenir et renforcer les liens de camaraderie et de solidarité entre ses membres ; leur permettre, par une action concertée, dans le cadre de la F.N.A.C.A., d'assurer la sauvegarde de leurs droits matériels et moraux et d'œuvrer en faveur de la paix, par la commémoration annuelle de l'anniversaire du cessez-le-feu ayant mis fin à la guerre en Algérie, le 19 mars 1962"
- **3 déclarations** avec "entretenir et renforcer les liens de camaraderie et de solidarité entre les anciens mobilisés en Algérie, Maroc et Tunisie ; leur permettre par une action concertée ; d'assurer la sauvegarde de leurs droits matériels et moraux et d'œuvrer en faveur de la paix".

Autres fédérations nationales (variantes non listées) :

- 2 déclarations avec "A.R.A.C." dans Titre
- 3 déclarations avec "A.M.M.A.C." dans Titre
- 2 déclarations avec "AMICALE DES ANCIENS COMBATTANTS D'AFRIQUE DU NORD" dans Titre
- 16 déclarations avec "AMICALE DES ANCIENS COMBATTANTS DE"
  + 27 variantes de "AMICALE DES ANCIENS COMBATTANTS" + MUTILES DE GUERRE / PRISONNIERS DE GUERRE / D' / DES / DU / ET VEUVES DE GUERRE / etc.
- 63 déclarations avec "ASSOCIATION DES ANCIENS COMBATTANTS" avec variantes (dont 22 avec "ASSOCIATION DES ANCIENS COMBATTANTS DE", etc.