L'apport du numérique aux sciences historiques: exemple d'une analyse computationnelle des archives Werner

Note méthodologique

Frédéric Clavert[1], ingénieur de recherche, LabEx *Écrire une Histoire Nouvelle de l'Europe* / IRICE (Paris Sorbonne)

Introduction

En 2007, après une série d'articles dans la *New Left Review*[2], Franco Moretti, spécialiste d'histoire de la littérature et professeur à Stanford, avance dans son livre *Graphs, Maps and Trees* la notion de lecture « distanciée » ou « distante » (*distant reading*)[3]. Constatant que l'histoire de la littérature ne concernait le plus souvent que l'histoire des « grands textes », il propose d'appliquer des méthodes computationnelles – l'informatique et la statistique appliquées à l'histoire de la littérature – pour intégrer dans cette histoire l'ensemble des textes produits au XVIII^e^ et XIX^e^ siècles ou, à tous le moins, plus que le faible pourcentage de la production littéraire de l'époque que représentent les « grands textes ». La première référence citée par Moretti est l'application des méthodes statistiques utilisées par l'école des Annales dans les années 1960 (*graphs*)[4], qu'il associe aux cartes des géographes (*maps*) et à la théorie de l'évolution (*trees*).

---

1 Cette recherche a été démarrée alors que l'auteur était encore membre du Centre Virtuel de la Connaissance sur l'Europe.

2 Franco Moretti, 'Graphs, Maps, Trees - 1', *New Left Review*, 24 (2003) <http://newleftreview.org/II/24/franco-moretti-graphs-maps-trees> [accessed 3 October 2012]; Franco Moretti, 'Graphs, Maps, Trees - 2', *New Left Review*, 26 (2004) <http://newleftreview.org/II/26/franco-moretti-graphs-maps-trees> [accessed 3 October 2012]; Franco Moretti, 'Graphs, Maps, Trees - 3', *New Left Review*, 28 (2004) <http://newleftreview.org/II/28/franco-moretti-graphs-maps-trees> [accessed 3 October 2012].

3 Franco Moretti, *Graphs, Maps, Trees: Abstract Models for Literary History* (Verso, 2007).

4 L'École des Annales a été pionnière dans l'utilisation de l'informatique en histoire, utilisant des méthodologies computationnelles quelques années après l'arrivée du premier supercalculateur (*mainframe*) en France. Voir François Furet and Adeline Daumard, 'Méthodes de l'Histoire sociale: les Archives notariales et la Mécanographie', *Annales ESC*, 14 (1959), 676–93; Paul Garelli and Jean-Claude Gardin, 'Étude Par Ordinateurs Des Établissements Assyriens En Cappadoce', *Annales ESC*, 16 (1961), 837–76. Dès cette époque, deux éléments dans l'usage de l'informatique en histoire se dégagent : la nécessité de gérer une abondance d'information d'une part, de croiser des données provenant de sources distinctes pour en faire émerger de nouvelles informations d'autre part.

La notion de lecture distanciée est emblématique des nouvelles méthodologies qui apparaissent à l'ère numérique et sont applicables aux sciences humaines et sociales. Dans le cas de l'histoire, elle implique de nouveaux modes de lecture des sources primaires, que nous avons explorés à plusieurs reprises[5]. En elle-même, la proposition de Moretti n'est pas si révolutionnaire : lors de recherches d'ampleur, tout chercheur a besoin de prendre du recul pour analyser ses sources, de voir ses archives comme un ensemble de documents reliés entre eux, prenant du sens les uns en relation avec les autres. Dans le cas du comité Werner, il est difficile d'en comprendre la logique si l'on ne considère pas le contexte monétaire des années 1960, c'est-à-dire le dérèglement progressif de Bretton Woods, et des années 1970 – la fin de ce système monétaire international et le grand changement de paradigme des politiques économiques soit le glissement du keynésianisme vers le monétarisme et la théorie de l'offre.

Ce qui est facteur de changement « disruptif[6] » chez Moretti est la possibilité d'appliquer ces méthodologies à des ensembles massifs de sources primaires, ce que l'on qualifierait aujourd'hui de *Big Data* appliqué à l'histoire[7]. La « révolution », pour user d'un terme galvaudé dans le monde numérique, vient de ces quantités massives d'information que l'ordinateur nous permet de traiter.

Nous allons dans ce chapitre essayer d'appliquer cette notion de lecture distanciée au corpus publié par le *Centre Virtuel de la Connaissance sur l'Europe* sur les activités du comité Werner, bien que nous ne soyons pas ici face à un corpus regroupant des données massives. Nous structurerons notre chapitre en trois sections : un exposé méthodologique, un exposé des résultats et une réflexion sur les limites de cette méthode.

Méthodologie

> « We know how to read primary sources. It is time to learn how not to read them[8] »

Nous allons utiliser ici un ensemble de méthodologies se rattachant à ce qui est appelé analyse ou fouille de texte (*text mining*). Nous les avons appliquées à un ensemble de textes tiré du Corpus publié par le CVCE.

---

5 Voir notamment Frédéric Clavert, 'Lecture Des Sources Historiennes À L'ère Numérique', *Frédéric Clavert*, 2012 <http://www.clavert.net/?p=1061> [accessed 15 March 2013].

6 La « disruption » est un anglicisme. Le terme désigne une technologie, peu remarquée à ses débuts car donnant dans un premier temps des résultats médiocres, qui finit par prendre la place des anciennes technologies au fur et à mesure de ses améliorations. Dans le cas présent, parler de méthodologie plus que de technologie serait plus pertinent.

7 Cette notion a été développée dans Patrick Manning, *Big Data in History*, 2013.

8 Inspiré d'une citation de Franco Moretti : « We know how to read books. It is time to learn how not to read them. » in Franco Moretti, *Distant Reading* (London: Verso, 2013). Location 794, édition Kindle.

Constitution et descriptif du corpus

Pour constituer notre corpus, nous avons tout d'abord téléchargé l'ensemble des textes disponibles en ligne au format PDF du Corpus Werner de la bibliothèque numérique consacrée à l'histoire de la construction européenne du CVCE, à l'aide de la ligne de commande UNIX *wget*. Le travail le plus important a ensuite été de classer ces textes.

Nous avons ainsi constitué un ensemble ayant diverses caractéristiques. Ce sont d'abord des textes appartenant directement aux archives du comité Werner. Nous n'avons analysé aucun document lié au contexte – les articles de presse, par exemple, n'ont pas été inclus -, ni aucune source liée à la désignation des membres du comité. Le rapport final lui-même est exclu. Le but est de se concentrer sur les débats qui ont eu lieu au sein du groupe Werner.

Nous avons ainsi constitué un ensemble (le corpus) contenant 77 textes rassemblant 580 548 occurrences (les mots) et 31 937 formes. Les formes sont des « types de mots » : le corpus est dit lemmatisé, c'est-à-dire que tout verbe conjugué est ramené à l'infinitif, les adjectifs au masculin singulier et les noms à leur forme singulière.

Le corpus compte 18 850 HAPAX, c'est-à-dire occurrences qui n'apparaissent qu'une seul fois dans le texte, ce qui correspond à 3,25 % des occurrences totales et 59,02 % des formes. Ces pourcentages sont élevés, nous y reviendrons dans la troisième section.

Méthodes appliquées

Une fois notre corpus constitué, nous avons appliqué un ensemble d'analyses ressortissant de la fouille de texte. Nous utilisons ici une approche statistique du texte. Chaque texte est divisé en segments de texte de quarante mots en moyenne, soit 16 147 au total.

Le nuage de mots (*wordcloud* ou *tagcloud*)

Le nuage de mots est l'analyse la plus simple que nous ayons effectuée. Elle donne à chaque forme lemmatisée une taille fondée sur le nombre d'occurrences de cette forme dans l'ensemble du corpus. Malgré certains pièges liés à ce type de visualisation[9], le nuage de mot peut être une bonne introduction dans ce type de démarches, mais elle ne sera jamais suffisante.

La Classification méthode GNEPA (anciennement Alceste)

La classification méthode GNEPA (anciennement Alceste, du nom du logiciel qui l'a introduite pour la première fois) procède à une classification hiérarchique descendante, permettant de dégager des profils de mots (des ensemble de mots se retrouvant souvent associés dans les mêmes segments de texte) que l'on peut ensuite essayer d'interpréter. D'un usage parfois délicat, elle laisse une grande marge de manœuvre au chercheur quant au nombre de profils (de thèmes) que l'on souhaite faire ressortir. Chaque profil, ou classe, peut être associé à un nuage de mots ou à une analyse de similitude (voir plus loin).

---

9 La proximité des formes entre elles dans le nuage, leur sens, leur couleur sont aléatoires donc ne sont pas interprétables mais peuvent brouiller le message de ce type de visualisation.

L'analyse de similitude

L'analyse de similitude s'intéresse aux liens existant entre les mots, une cooccurrence dans un segment de texte formant un lien en deux mots. L'analyse de similitude permet de réaliser une visualisation du réseau de mots que forme le corpus.

Logiciels utilisés

Chaque logiciel ayant ses propres particularités, nous préférons ici préciser les logiciels que nous avons utilisés. L'application principale employée est IRaMuTeQ[10] dans sa version 0.6 alpha 3, logiciel libre développé par Pierre Ratinaud (Toulouse) et qui s'inspire de l'historique (et payant) Alceste. IRaMuTeQ est en fait une interface graphique spécialisée dans l'analyse de texte du programme statistique libre R[11], ici dans sa version 2.15.3.

Les résultats des analyses sont exposés dans la section suivante.

Les grands sujets de discussion des membres du comité Werner

10 http://www.iramuteq.org/

11 http://www.r-project.org/

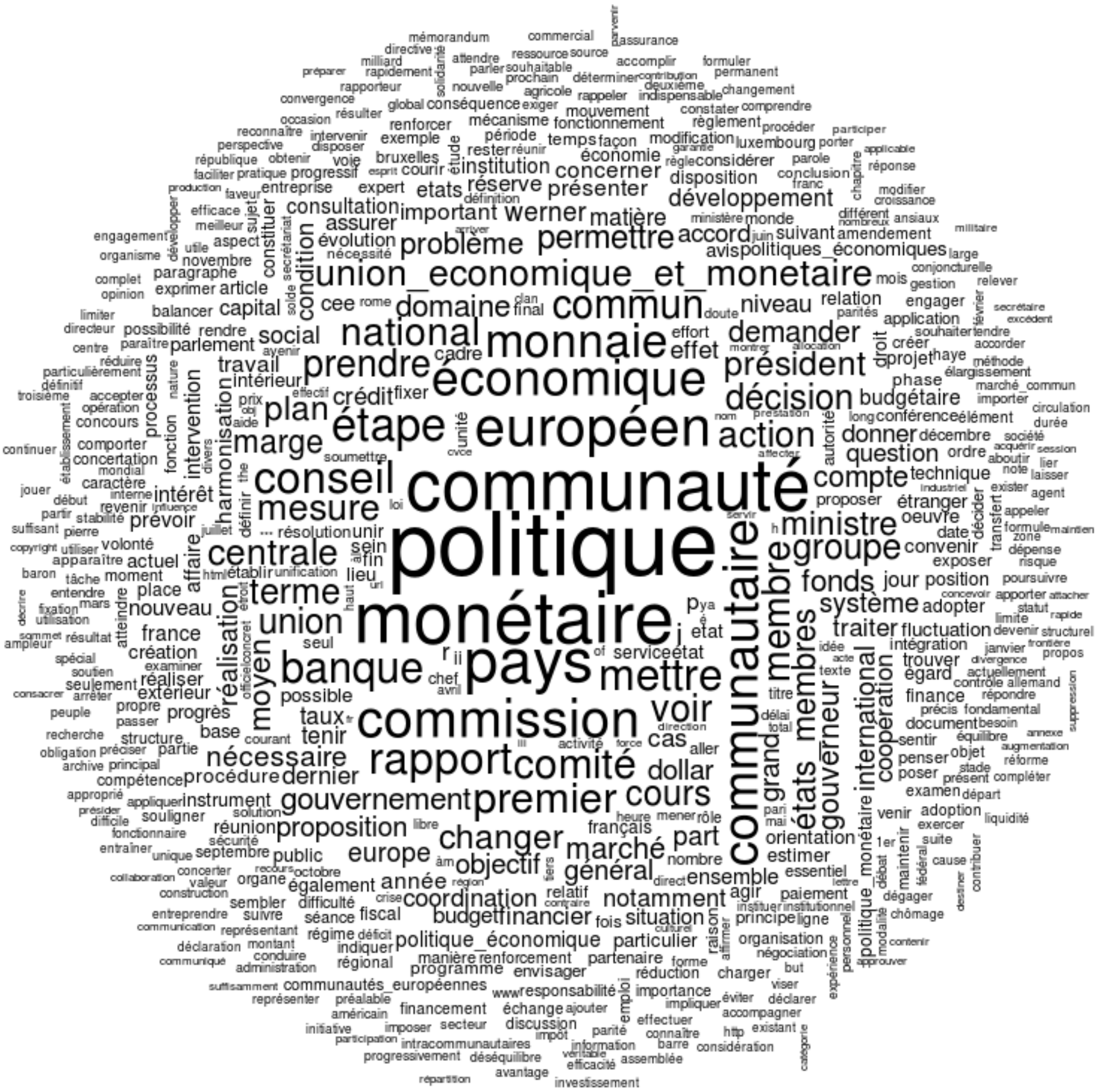

Figure 1 - Nuage de mots de l'ensemble du corpus

Le nuage de mots de l'ensemble du corpus (Figure) nous permet d'avoir un premier aperçu du contenu du corpus constitué. Le terme « politique » est très clairement le plus utilisé. Logiquement les formes qui suivent sont des adjectifs en rapport aux missions du comité Werner (*économique*, *monétaire*), aux institutions européennes (*commission*, *conseil* mais aussi *banque, centrale* probablement lié à la demande d'avis effectuée auprès du Comité des gouverneurs mais aussi aux partage des responsabilités entre diverses institutions, dont les banques centrales, en cas d'Union économique et monétaire), ou des termes liés aux pays membres (*pays, national*). Une grande partie du vocabulaire ainsi montré tient également aux membres du comité (*président*), au contenu des discussions (*union, étape, etc).*

Si le terme « politique » est le plus utilisé, c'est à cause de la manière dont l'action de la Communauté économique européenne définit son action, c'est-à-dire par ses politiques (économique, agricole, budgétaire, monétaire) notamment en relation et en interaction avec celles des États membres.

Comment ces termes sont-ils reliés ensemble ? L'analyse de similitude permet de le préciser :

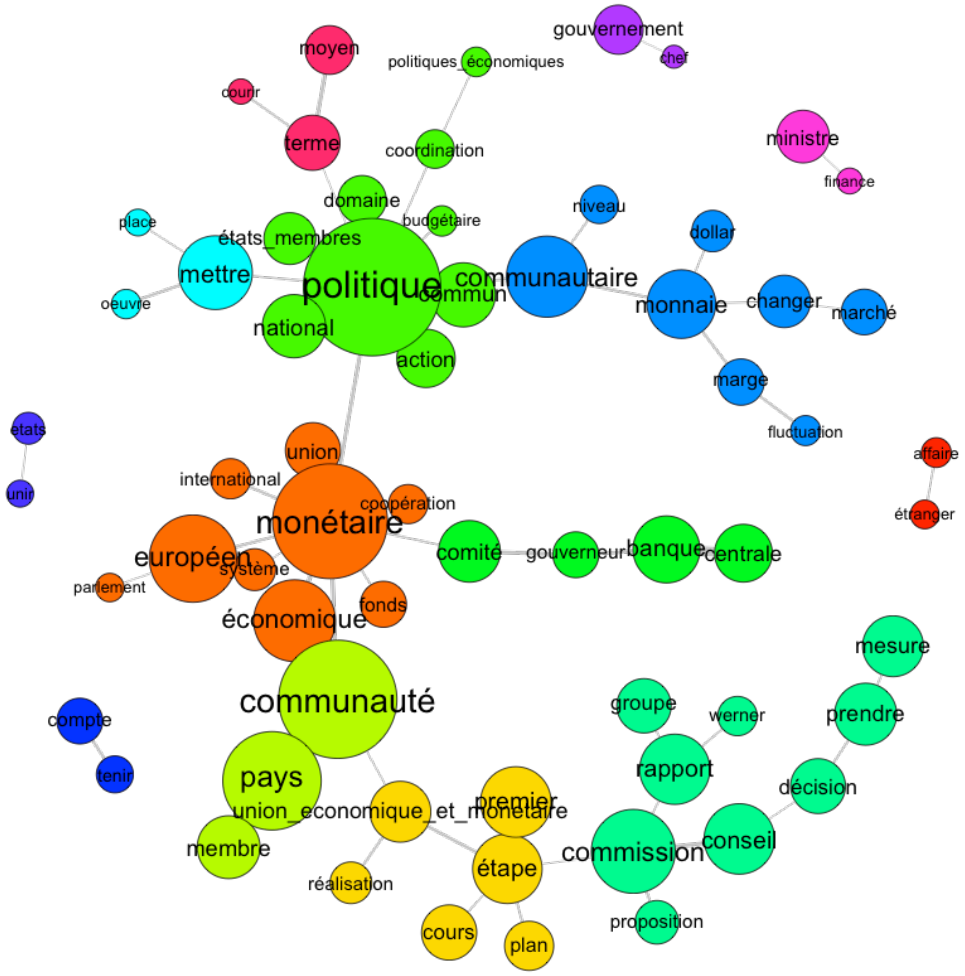

Figure 2 - Graphe de l'analyse de similitude

L'analyse de similitude (Figure) nous permet de voir les relations entre les mots au sein du texte. Sans entrer dans les détails, on peut déduire de cette analyse que le corpus se structure autour des termes « politique », « monétaire » et « communauté » : le sujet essentiel de discussion est ainsi la politique monétaire de la Communauté, ce qui correspond à la mission donnée au Comité. Les États membres semblent finalement un sujet assez périphérique de cette ossature essentielle du texte.

Pour affiner cette analyse, nous pouvons nous pencher sur la classification par la méthode Alceste (Figure). Le sujet le moins relié aux autres est celui relatif à la forme des documents (classe 4) qui touche au titre de Pierre Werner (président du Comité) et aux relations avec le Conseil et la Commission, avec certains ministres. Cette classe regroupe 26,3% des segments de texte. Le respect des formes, pour ce type de comité, est ainsi important.

La classe qui regroupe le plus grand nombre de segments de texte est la première, qui rassemble un vocabulaire touchant à la politique communautaire dans le domaine monétaire et économique. C'est le cœur des discussions au sein du comité Werner. Elle est reliée directement à la classe 2, qui ne touche que 16,3% des segments de texte et dont les termes semblent se référer à des questions très techniques : les marges de fluctuation des monnaies européennes, notamment, y compris en relation au dollar.

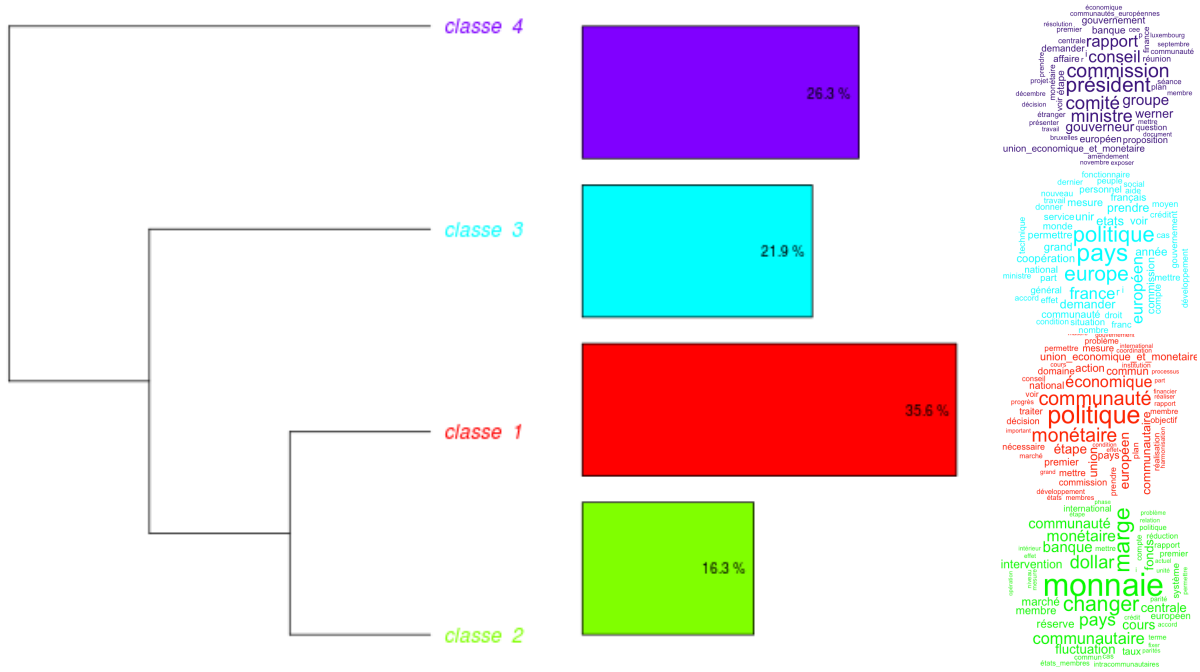

Figure 3 - Dendogramme et nuages de mots correspondant aux quatre classes dégagées par l'analyse Alceste.

La classe n° 3 (21,9% des segments de texte) touche plus aux relations entre pays membres et Europe, y compris dans le domaine monétaire.

Regardons désormais d'un peu plus près la classe 1, en procédant à une analyse de similitude (Figure 4 - Classification de la première classe) ainsi qu'une classification méthode Alceste (Figure 5 - Analyse de similitude de la classe 1) spécifiques à ce profil.

La seconde montre que la classe 1 est divisée en trois plus grands thèmes : les discussions sur le rapport lui-même et, surtout, sur le contenu de la première étape décrite par le rapport, le contexte d'une mise en place d'une Union économique et monétaire (y compris le contexte internationale et la notion de réduction des déséquilibres – monétaires, économiques, politiques) et, enfin, la répartition future des compétences entre différents organes, dont les banques centrales et leurs gouverneurs.

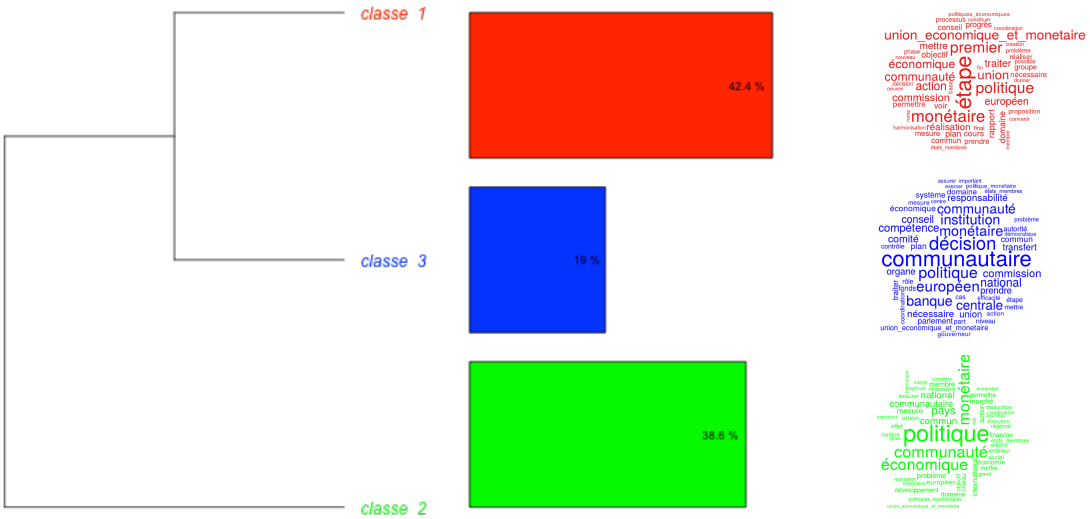

Figure 4 - Classification de la première classe

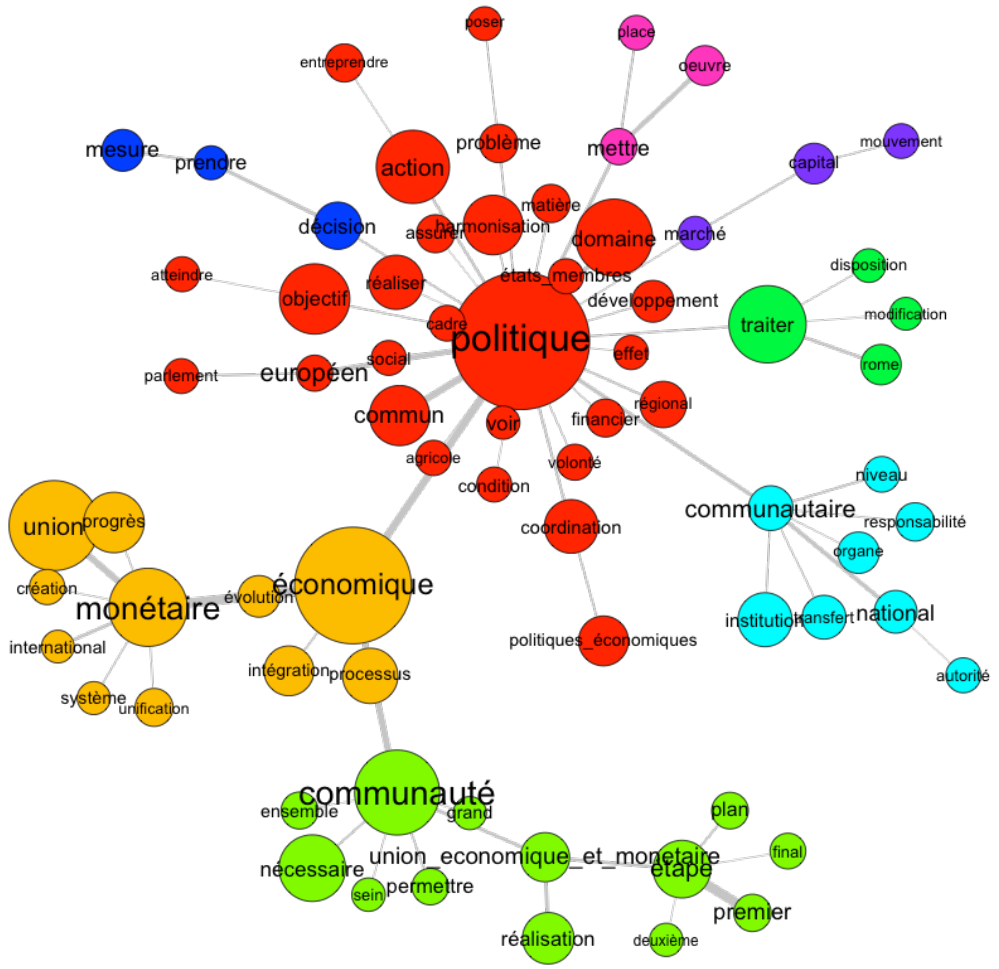

Figure 5 - Analyse de similitude de la classe 1

Dans l'ensemble, l'image donnée du corpus par son analyse computationnelle est conforme à l'état des connaissances sur le plan Werner. Ce qui manque, malheureusement, est une analyse des controverses. Toutefois, en l'état des comptes-rendus des séances du comité Werner, elle n'a pas été possible.

Limites méthodologiques

Nous allons ici analyser trois grandes limites méthodologiques que notre traitement informatique des archives Werner illustre très bien : la sélection, la langue, le processus de mise en données des archives.

Le biais de sélection

Le premier biais de cette analyse, difficile à quantifier, est lié à notre mode de sélection des archives, ou, plutôt, à l'empilement de trois types de sélection des textes utilisés.

La première sélection est celle de la famille de Pierre Werner elle-même, qui, de fait, a fait office d'archiviste[12]. Comme tout archiviste, des choix de conservation ont été opérés. En archivistique, les critères de sélection, explicites, sont effectués à plusieurs stades et deviennent irréversibles quand les documents acquièrent le statut d'archives définitives. Dans le cas des archives Werner, il n'est pas possible de connaître les modalités de la sélection, ce qui est souvent (et normalement) le cas des archives familiales.

La seconde sélection est celle qui a été effectuée par l'équipe du CVCE, autour de la chercheuse qui a mené le projet de numérisation et de publication de ces archives, Elena Danescu. Ayant fait partie à ce moment-là de l'équipe du CVCE, les critères de sélection nous sont plus clairs. Le corpus publié par le CVCE avait pour but de cerner les travaux du comité. Les documents publiés ont été sélectionnés en fonction de ce critère essentiel. Suivi par un « comité de pilotage » constitué de chercheurs aguerris en histoire économique et monétaire[13], les papiers Werner ont été complétés par de la documentation en provenance d'autres centres d'archives. Le corpus du CVCE compte au final de nombreux documents expliquant le choix de Werner comme président du comité, expliquant le contexte de la création du comité et de sa mise en place, expliquant les conditions de ses travaux, de son échec immédiat puis de sa postérité.

Le troisième niveau de sélection est celui opéré par l'auteur de ce chapitre. Nous avons souhaité nous concentrer, comme nous l'avons expliqué plus haut, sur les documents internes du Comité, relatant les débats entre ses membres. En effet, c'est ce point – les rapports de force au sein du comité – qui nous semble être le grand apport scientifique de la publication des archives familiales de Pierre Werner.

À ce biais, assumé, de sélection, qui, finalement, n'est pas propre à la nature numérique du corpus Werner mais plutôt à toute source primaire éditée, s'ajoute deux autres biais, plus en relation à ce type d'analyse computationnelle.

---

12 Pour avoir une idée sur la constitution du corpus, son contenu et lire une première analyse du rapport Werner à sa lumière, se reporter aux publications d'Elena Danescu sur et intégrées au corpus lui-même : Source: DANESCU, Elena Rodica. *Une relecture du rapport Werner du 8 octobre 1970 à la lumière des archives familiales Pierre Werner - étude approfondie (version intégrale)*. Sanem: CVCE, 2012. Disponible à l'adresse: www.cvce.eu.

13 Sylvain Schirmann (Université de Strasbourg, ancien directeur de thèse de l'auteur de l'article), René Leboutte (Université du Luxembourg) et Ivo Maes (Banque nationale de Belgique).

Le biais linguistique

La plupart des logiciels de fouille de texte, et IRaMuTeQ n'est pas une exception, sont adaptés au traitement de plusieurs langues, mais ne peuvent accepter que des corpus monolingues[14] : il est possible ainsi d'analyser des corpus en Allemand, Français, Anglais,… mais ces corpus ne peuvent comporter qu'une seule langue. L'ensemble de textes que nous avons soumis au logiciel est ainsi uniquement francophone. La conséquence est une sous-représentation du point de vue allemand. Une perspective comparatiste aurait pu être menée, en procédant à la création d'un corpus allemand. Ce dernier, toutefois, était trop restreint pour donner des résultats par une lecture computationnelle : la simple lecture humaine reste plus efficace dans ce cas.

Toutefois, ce biais linguistique peut être compensé, justement, par cette lecture humaine des textes en allemand. Le principal défaut de notre méthode, dans le cas précis, est lié au processus de mise en données du corpus Werner.

Le biais de la reconnaissance de texte et l'importance du processus de mise en données des sources primaires

La « mise en données » (*datafication*) est un processus décrit ainsi : « To datafy a phenomenon is to put it in a quantified format it can be tabulated and analysed »[15]. Appliquée à l'histoire[16], on peut estimer qu'elle regroupe l'ensemble des étapes nécessaires pour passer d'archives « analogiques » en données numériques structurées, ce qui inclut notamment la numérisation, la reconnaissance du texte (y compris, éventuellement, la reconnaissance d'entités : les noms propres, les lieux, les éléments de temporalités) et sa correction, la mise en place de métadonnées fiables.

Or, dans le cas du corpus Werner, le processus de mise en données minore la fiabilité des résultats obtenus et, cela, en raison du mode de reconnaissance de texte employé. De nombreux documents utilisés ici portent une mention avertissant que le résultat de l'OCR[17] n'a pas été corrigé. Cet avertissement explique un élément cité plus haut dans ce chapitre, à savoir le taux important d'HAPAX. À l'origine, ce taux était encore plus élevé et rendait toute exploitation computationnelle du corpus Werner impossible. Nous avons pu toutefois corriger ces défauts pour partie, grâce à un algorithme écrit par Sascha Kaufmann, alors membre du CVCE. Mais cette correction de la reconnaissance de caractère ne permet pas de savoir quel est, *in fine*, le taux de réussite de reconnaissance de texte de notre corpus. L'autre limite induite par la mise en données des archives Werner, plus à la marge, est qu'elle n'inclut pas les notes manuscrites de Pierre Werner :

---

14 Pour avoir une petite idée des logiciels disponibles, voir : Étienne Ollion, "Analyse quantitative de contenus 2.0," *Data Sciences Sociales*, 26-Jul-2014. [Online]. Available: http://data.hypotheses.org/948. [Accessed: 30-Sep-2014].

15 Viktor Mayer-Schönberger and Kenneth Cukier, *Big Data: A Revolution That Will Transform How We Live, Work, and Think* (Boston: Houghton Mifflin Harcourt, 2013). P. 78.

16 Pour plus d'éléments, voir : Frédéric Clavert, 'Mise En Données Du Monde, Mise En Données de L'histoire?', *Frédéric Clavert*, 2013 <http://www.clavert.net/mise-en-donnees-du-monde-mise-en-donnees-de-lhistoire/> [accessed 10 October 2013].

17 *Optical Caracter Recognition*.

si elles sont disponibles sous forme d'images, leur texte n'a pas été soumis à un processus d'OCR.

L'enjeu pour les sciences humaines et sociales que pose la reconnaissance de texte dans le processus de mise en données des archives n'est pas nouveau. Le cas du corpus Werner n'en est ici qu'une illustration parmi d'autres. Des historiens, comme Tim Hitchcock[18] ou Ian Milligan[19], se sont penchés sur cette problématique. Comme le rappelle Hitchcock, même un taux de reconnaissance de texte très élevé (99%) peut, si les erreurs portent sur un simple mot constitutif de la recherche par un chercheur, biaiser l'ensemble des résultats d'une recherche. Ainsi cette analyse du corpus Werner pose-t-elle les enjeux essentiels de l'application des méthodologies computationnelles en histoire.

Conclusions

Malgré les limites méthodologiques exposées dans la dernière section, l'image donnée par cette analyse du corpus Werner semble assez fidèle au contenu des documents, en tout les cas aux yeux de ceux qui ont « humainement » lu les archives. Il manquerait une analyse des controverses, néanmoins, pour mieux mettre en valeurs les oppositions et affinités entre membres du groupe Werner, et, notamment, pour savoir si la division entre deux groupes (« monétaristes » ou « économistes ») est aussi nette qu'on ne pourrait le préjuger *a priori*.

Les éléments méthodologiques doivent être pleinement explicitées, afin de mieux cerner les limites de notre analyse. D'une certaine manière, et comme le rappelle, d'ailleurs, Peter Haber dans son *Digital Past*[20], l'ère numérique doit nous inciter à revisiter les fondements de notre méthodologie d'historiens, et, surtout, à les revaloriser.

Enfin, remarque la plus importante, nous devons attacher la plus grande rigueur au processus de mise en données (du côté des professionnels de la numérisation et de l'archivistique) des sources primaires et à son analyse (du côté des chercheurs) si nous souhaitons que les investissements aujourd'hui fournis pour la numérisation de notre monde nous permettent d'aller au-delà d'une lecture humaine de la reproduction d'archives sur notre écran. La numérisation doit servir aux chercheurs, les aider à aller

---

18 Tim Hitchcock, 'Academic History Writing and Its Disconnects', *Journal of Digital Humanities*, Winter 2011 <http://journalofdigitalhumanities.org/1-1/academic-history-writing-and-its-disconnects-by-tim-hitchcock/> [accessed 27 June 2012]. Plus récemment dans le cadre d'un colloque organisé par le CVCE : Tim Hitchcock, 'Big Data for Dead People: Digital Readings and the Conundrums of Positivism', *Historyonics*, 2013 <http://historyonics.blogspot.co.uk/2013/12/big-data-for-dead-people-digital.html> [accessed 17 December 2013].

19 Ian Milligan, 'Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997–2010', *Canadian Historical Review*, 94 (2013), 540–69 <http://dx.doi.org/10.3138/chr.694>.

20 Peter Haber, *Digital Past. Geschichtswissenschaften Im Digitalen Zeitalter* (München: Oldenbourg Wissenschaftsverlag, 2011).

au-delà de leur lecture humaine traditionnelle, à articuler lecture proche et lecture distante, pour à terme améliorer la « fabrique de l'Histoire »[21].

Le corpus Werner n'est pas une exception. Nombreux sont les corpus qui, numérisés avec les moyens technologiques disponibles au moment de leur mise en données, ont à la fois les mêmes forces et faiblesses. Il apparaît ainsi nécessaire que les institutions qui pilotent des projets de numérisation mettent au point une méthodologie de mise en données la plus rigoureuse possible au vu des moyens techniques qui sont à leur disposition et n'hésitent pas à re-numériser ou à revenir sur tout ou partie de la mise en données quand ces moyens évoluent.

21 Pour employer l'excellent titre de la non moins très bonne émission d'Emmanuel Laurentin sur France Culture.